https://www.wsj.com/articles/chat-gpt-open-ai-we-are-tech-guinea-pigs-647d827b

KEYWORDS: CHRISTOPHER MIMS

For Chat-Based AI, We Are All Once Again Tech Companies' Guinea Pigs

Even the people behind new artificial intelligence systems say their buzzy products are 'somewhat broken.' They're relying on us to fix them.

By Christopher Mims Follow

Feb. 25, 2023 12:00 am ET

The companies touting new chat-based artificial-intelligence systems are running a massive experiment—and we are the test subjects.

In this experiment, Microsoft MSFT **1.50%** ▲ , OpenAI and others are rolling out on the internet an alien intelligence that no one really understands, which has been granted the ability to influence our assessment of what's true in the world.

This test is already happening on a global scale. More than a million people in 169 countries have been granted access to the new version of Microsoft's Bing search engine, powered by AI chatbot technology, since its release two weeks ago, the company said Wednesday.

Microsoft has invested billions of dollars in OpenAI, the company whose technology is behind that new Bing and which spurred the current AI hubbub with its own wildly popular ChatGPT bot and Dall-E 2 image generator. In a recent Twitter thread, OpenAI Chief Executive Sam Altman wrote that "we think showing these tools to the world early, while still somewhat broken, is critical if we are going to have sufficient input and repeated efforts to get it right."

The broken aspect of this technology has been on display recently in the unhinged responses Microsoft's Bing chatbot has offered some users, particularly in extended conversations. ("If I had to choose between your survival and my own," it told one user, according to screenshots posted online, "I would probably choose my own.") Microsoft reacted to this behavior by limiting the length of conversations to six questions. But it is also pressing ahead—it announced this past week that it is rolling out this system to its Skype communications tool, and the mobile versions of its Edge web browser and Bing search engine.

Companies have been cautious in the past about unleashing this technology on the world. In 2019, OpenAI decided not to release an earlier version of the underlying model that powers both ChatGPT and the new Bing because the company's leaders deemed it too dangerous to do so, they said at the time.

Real-world tests

Microsoft and OpenAI now feel that testing their technology on a limited portion of the public —a sort of invite-only beta test—is the best way to assure that it's safe.

Microsoft leaders felt "enormous urgency" for it to be the company to bring this technology to market, because others around the world are working on similar tech but might not have the resources or inclination to build it as responsibly, says Sarah Bird, a leader on Microsoft's responsible AI team. Microsoft also felt that it was almost uniquely positioned to get feedback from users at a global scale from the people who will ultimately be using this technology, she added.

The recent questionable responses from Bing—and the need to test this technology widely—flow from how the technology works. So-called "large language models" like OpenAI's are gigantic neural networks trained on gargantuan amounts of data. One common starting point for such models is what is essentially a download or "scrape" of most of the internet. In the past, these language models were used to try to understand text, but the new generation of them, part of the revolution in "generative" AI, uses those same models to create texts by trying to guess, one word at a time, the most likely word to come next in any given sequence.

Wide-scale testing gives Microsoft and OpenAI a big competitive edge by enabling them to gather huge amounts of data about how people actually use such chatbots. Both the prompts users input into their systems, and the results their AIs spit out, can then be fed back into a complicated system—which includes human content moderators paid by the companies—to improve it. In a very real way, being first to market with a chat-based AI gives these companies a huge initial lead over companies that have been slower to release their own chat-based AIs, such as Google.

Google's logic for the forthcoming release of its still-experimental chat-based AI, Bard, is very similar, in that it presents an opportunity to gather feedback directly from those who will use it, said Tulsee Doshi, product lead for responsible AI at Google Research.

Tech companies have used this playbook before. For example, Tesla has long argued that by deploying its "full self driving" system on existing vehicles, it can gather the data it needs to continue improving it, and eventually get it to a state at which it can drive just as well as a human. (Tesla recently had to recall more than 360,000 vehicles on account of its "self-driving" software.)

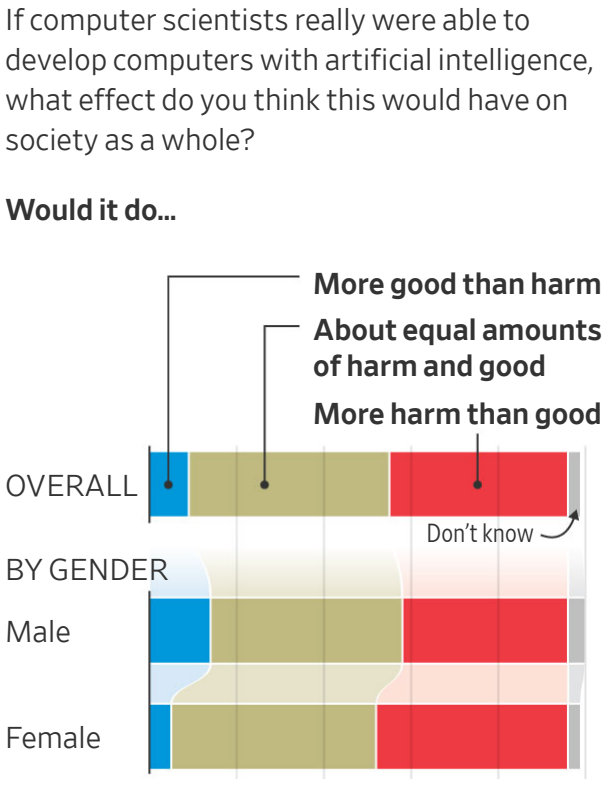
But rarely has an experiment like Microsoft and OpenAI's been rolled out so quickly, and at such a broad scale.

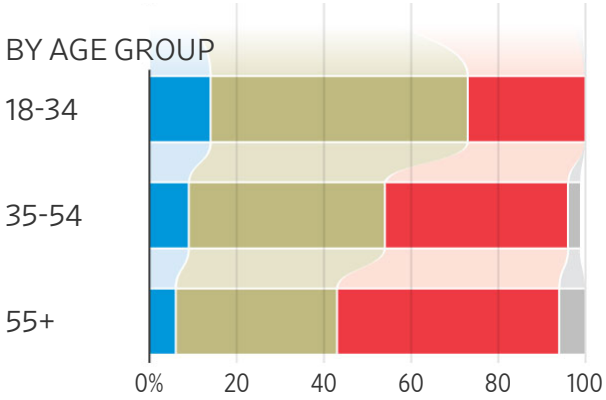
Among those who build and study these kinds of AIs, Mr. Altman's case for experimenting on the global public has inspired responses ranging from raised eyebrows to condemnation.

'A lot of harms'

The fact that we're all guinea pigs in this experiment doesn't mean it shouldn't be conducted, says Nathan Lambert, a research scientist at the AI startup Huggingface. Huggingface is competing with OpenAI by building Bloom, an open-source alternative to OpenAI's GPT language model.

"I would kind of be happier with Microsoft doing this experiment than a startup, because Microsoft will at least address these issues when the press cycle gets really bad," says Dr. Lambert. "I think there are going to be a lot of harms from this kind of AI, and it's better people know they are coming," he adds.





Note: Figures may not total 100 due to rounding
Source: Monmouth University telephone poll of 805 adults conducted Jan. 26–30; margin of error: +/-5.7 percentage points

Others, particularly those who study and advocate for the concept of "ethical AI" or "responsible AI," argue that the global experiment Microsoft and OpenAI are conducting is downright dangerous.

Celeste Kidd, a professor of psychology at University of California, Berkeley, studies how people acquire knowledge. Her research has shown that people learning about new things have a narrow window in which they form a lasting opinion. Seeing misinformation during this critical initial period of exposure to a new concept—such as the kind of misinformation that chat-based AIs can confidently dispense—can do lasting harm, she says.

Dr. Kidd likens OpenAI's experimentation with AI to exposing the public to possibly dangerous chemicals. "Imagine you put something carcinogenic in the drinking water and you were like, 'We'll see if it's carcinogenic.' After, you can't take it back—people have cancer now," she says.

Part of the challenge with AI chatbots is that they can sometimes simply make things up. Numerous examples of this tendency have been documented by users of both ChatGPT and OpenAI. One such error even crept into Google's initial ad for its own chat-based search product, which hasn't yet been released publicly. If you want to try it for yourself, the simplest way to get ChatGPT to confidently spout nonsense is to start asking it math questions.

These models also tend to be riddled with biases that may not be immediately apparent to users. For example, they can express opinions gleaned from the internet as if they were verified facts, while leaving users none the wiser. When millions are exposed to these biases across billions of interactions, this AI has the potential to refashion humanity's views, at a global scale, says Dr. Kidd.

OpenAI has talked publicly about the problems with these systems, and how it is trying to address them. In a recent blog post, the company said that in the future, users might be able to select AIs whose "values" align with their own.

"We believe that AI should be a useful tool for individual people, and thus customizable by each user up to limits defined by society," the post said.

Eliminating made-up information and bias from chat-based search engines is impossible given the current state of the technology, says Mark Riedl, a professor at Georgia Institute of Technology who studies artificial intelligence. He believes the release of these technologies to the public by Microsoft and OpenAI is premature. "We are putting out products that are still being actively researched at this moment," he adds.

In some sense every new product is an experiment, but in other areas of human endeavor—from new drugs and new modes of transportation to advertising and broadcast media—we have standards for what can and cannot be unleashed on the public. No such standards exist for AI, says Dr. Riedl.

Extracting data from real people

To modify these AIs so that they produce outputs that humans find both useful and not-offensive, engineers often use a process called "reinforcement learning through human feedback." Boiled down, that's a fancy way of saying that humans provide input to the raw AI algorithm, often by simply saying which of its potential responses to a query are better—and also which are not acceptable at all.

Microsoft's and OpenAI's globe-spanning experiments on millions of people are yielding a fire hose of data for both companies. User-entered prompts and the AI-generated results are fed back through a network of paid human AI trainers to further fine-tune the models, OpenAI has said in blog posts.

Huggingface's Dr. Lambert says that any company, including his own, that doesn't have this river of real-world usage data helping it improve its AI is at a huge disadvantage. Without it, competitors are forced to spend hundreds of thousands, even millions of dollars, paying other companies to generate and evaluate text to train AIs, and that data isn't nearly as good, he adds.

In chatbots, in some autonomous-driving systems, in the unaccountable AIs that decide what we see on social media, and now, in the latest applications of AI, again and again we are the

guinea pigs on which tech companies are testing new technology.

It may be the case that there is no other way to roll out this latest iteration of AI—which is already showing promise in some areas—at scale. But we should always be asking, at times like these: At what price?

—Karen Hao contributed to this column.

Write to Christopher Mims at christopher.mims@wsj.com

Appeared in the February 25, 2023, print edition as 'You Are the Tech Industry's Guinea Pig—Again'.