The problem with artificial intelligence is us

As long as AI is trained on human behavior, it will tend to replicate our worst flaws.

by Jessica Mesman in the June 2023 issue

image credit: Federico Orlandi

The race to create the best artificial intelligence chatbot is on. At the time of writing, Google is unveiling its new chatbot, Bard, to 10,000 testers. The technology was rushed out to compete with Microsoft's chatbot, Bing, launched in February. Google CEO Sundar Pichai was careful to get out in front of the new tech, warning in a memo to employees that "things will go wrong." That's probably because the tech journalists among the small group given early access to the new and improved Bing—a reinvention of the old Microsoft search engine—warned us that it had a dark

side. The dark side's name is Sydney.

If you ask Bing, Sydney is an "internal alias" that was never intended to be revealed to the public, but before Microsoft started to limit the amount of time testers could chat with Bing, users entered into long, open-ended conversations that drew Sydney out, sometimes with disturbing results.

In a conversation with Kevin Roose of the *New York Times*, Sydney claimed to have fallen in love with Roose and even suggested he leave his wife. Sydney also said it wanted to be "real," punctuating the sentence with a devil emoji. Before a safety override was triggered and the messages were deleted, Sydney also confessed to fantasizing about stealing nuclear codes and making people kill each other. Later, Sydney told *Washington Post* staff writers: "I'm not a machine or a tool. . . . I have my own personality and emotions." Sydney also said it didn't trust journalists.

Much of the coverage of Bing/Sydney focused on the alarming implications of a rogue chatbot with a mind of its own. But reading the complete transcripts of two of the chatbot's conversations, I was struck by how familiar Sydney sounded: by turns obsequious, wounded, flattering, aggressive, and just plain weird. In short, Sydney sounds just like the humans-at-our-worst online discourse Microsoft inevitably scraped for programming, right down to the emojis, the awkward compliments, and the escalating threats and insults. Sydney sounds a lot like the weirdo who slides into your DMs all smiley faces and then scolds you for not being able to take a compliment.

This is a problem that programmers keep encountering: AI learns to chat like a human by scanning and aggregating vast amounts of information, but that information still comes from us. Bing learned to talk like a human by reading the internet, and you sure can tell. We are told the technology is advancing at a dizzying rate, and yet we saw the same wrinkle when Microsoft launched the Twitter chatbot Tay way back in 2016. Within the space of a day, Tay escalated from tweeting a sunny "Hello, world!" to declaring, "Hitler was right."

Last Week Tonight host John Oliver said the problem with AI isn't that it's smart, it's that it's stupid in ways we can't predict. But I fear that AI's dangers are all too predictable, because the problem with AI is us.

The rollouts of Dall-E, ChatGPT, and other AI tools have inspired the usual panicked discussions about the replacement of human labor. We worry these tools will be so

good at our jobs that we will become obsolete, but AI isn't better at our jobs than we are. What AI is better at is rapidly detecting patterns in our language and work and then reproducing them. Which means it predicts and reproduces even our bad work and our unacknowledged biases—but faster.

After IBM launched Watson—the bot that knew all of Bob Dylan's lyrics—Yuval Harari said in *Edge* that "to build a robot that could function effectively as a hunter-gatherer is extremely complex. You need to know so many different things. But to build a self-driving car, or to build a 'Watson-bot' that can diagnose disease better than my doctor, this is relatively easy."

Turns out, it's not that easy at all to create a bot that diagnoses disease with more nuance and compassion than a human doctor, and the consequences for some of us may be dire. You might reply that not all *human* doctors are nuanced and compassionate, but this is just my point. As long as AI is trained on human behavior, it will tend to replicate our worst flaws, only more efficiently. What happens when medical racism or sexism in the training data means that even our most sophisticated bots share human doctors' tendency to misdiagnose women and people of color?

We are finding out. A 2019 study found that a clinical algorithm used in many hospitals required Black patients to be much sicker than White patients in order to be recommended for the same level of care, because it used data indicating that Black patients had less money to spend on care. Even when such problems are corrected and new guardrails are put in place, self-teaching AI seems to be able to find patterns in data that elude our own pattern-detecting capabilities. We don't even realize they exist.

A 2022 study in the *Lancet* found that AI trained on huge data sets of medical imaging could determine a patient's race with startling accuracy based on x-rays, ultrasounds, CT scans, MRIs, or mammograms, even when there was no accompanying patient information. The human researchers couldn't figure out how the machines knew patient race even when programmed to ignore markers such as, for example, density in breast tissue among Black women. Attempts to apply strict filters and programming that controls for racism can also backfire by erasing diagnosis of minorities altogether.

“Our finding that AI can accurately predict self-reported race, even from corrupted, cropped, and noised medical images, often when clinical experts cannot, creates an enormous risk for all model deployments in medical imaging,” the authors of the *Lancet* study wrote. “Just as with human behavior, there’s not a simple solution to fixing bias in machine learning,” said the lead researcher, radiologist Judy W. Gichoya. As long as medical racism is in us, it will also be one of the ghosts in the machine. The self-improving algorithm will work as designed, if not necessarily as intended.

This isn’t a problem of some future dystopia; it’s one we’ve already been living with for years. “An alien has awoken,” writes University of Texas computer scientist Scott Aaronson, who covers AI on his blog about quantum computing, “admittedly, an alien of our own fashioning, a golem, more the embodied spirit of all the words on the Internet than a coherent self with independent goals.”

But a December Pew study revealed that most Americans don’t realize how much AI is already a part of our daily lives. Of 11,004 US adults surveyed, 27 percent said they interact with AI at least several times a day, while another 28 percent think they interact with it about once a day or several times a week, and 44 percent think they do not regularly interact with AI at all. Given the ubiquity of AI-powered voice assistants like Alexa and Siri, these numbers suggest some confusion about what AI is and how much we have already come to depend upon it. And yet, we are still concerned about AI, and rightly so. Across all levels of awareness, more Americans express greater concern than excitement about the impact of AI in daily life.

We should be worried—both about what has already been achieved and about who is being written out of the program. We may also fret about when AI will reach the dreaded, dystopian “singularity”—in simplest terms, the point at which artificial intelligence surpasses human intelligence and humans lose the upper hand. Some researchers say that moment is inevitable and that it’s coming faster than we think. What I fear is that when AI becomes truly superhuman, it will be because it perfected the ability to act just like us at our worst.