Bruce Schneier
Harvard University

Trustworthy AI Means Public AI

Back in 1998, Sergey Brin and Larry Page introduced the Google search engine in an academic paper that questioned the ad-based business model of the time. They wrote: "We believe the issue of advertising causes enough mixed incentives that it is crucial to have a competitive search engine that is transparent and in the academic realm." Although they didn't use the word, their argument was that a search engine that could be paid to return particular URLs is fundamentally less trustworthy. "Advertising income often provides an incentive to provide poor quality search results."

We all know what happened next. Google eventually sold ads: first in dedicated boxes that made it obvious that the links were paid for, and then slowly integrated into the actual search results until paid links were seamlessly integrated into the "real" search results.

It's a story that's been repeated many times. Your Facebook and Instagram feeds are filled with "sponsored posts." An Amazon search returns pages of products whose sellers paid for placement. Buried behind an obscure link, the travel search engine Kayak discloses that "some results have been promoted based on their revenue potential for us."

This is how the Internet works. Companies spy on us as we use their services. Data brokers buy that information from smaller companies, and assemble detailed dossiers on us. They then sell that information back to other companies, who combine it with data they collect in order to manipulate our behavior to serve their interests—at the expense of our own.

And even if they don't manipulate us directly—even if we pay for the products and services—they spy on us. Our televisions spy on us. Our smartphones spy on us, as do our appliances, our cars, and everything else with a plug or battery. Surveillance is the business model of the Internet. And the use of manipulative interfaces is so prevalent that it has its own name: dark patterns.

We use all of these services as if they were our agents. In fact, they are double agents, secretly working for their corporate owners. We trust them, but they are not trustworthy.

We should not expect the current crop of generative AI systems to be any different. And the results will be much worse, for two reasons.

The first is that these AIs will be more relational. We will be conversing with these systems, using natural language. As such, we will naturally ascribe human-like characteristics to them. And we will treat them as trusted assistants and intimate friends.

This relational nature makes it easier for double agents to do their work. Did your chatbot recommend a particular airline or hotel because it's truly the best deal, given your particular set of needs, or because the AI company got a kickback from those providers? When you asked it to explain a political issue, did it bias that explanation towards the company's position? Or in a way that benefits whichever political party gave it the most money?

The second reason to be concerned is that these AIs will be more intimate. One of the promises of generative AI is a personal digital assistant: acting as your advocate with others, and as an intimate butler with you. This requires an intimacy greater than your search engine, email provider, cloud storage system, or smartphone. You're going to want it with you 24 × 7, constantly recording and training on everything you do, so it can most effectively work in your best interest.

And it will help you in many ways. It will notice your moods and know what to suggest. It will anticipate your needs and work to satisfy them. It will be your therapist and life coach.

You will want to trust it. Its interface will make it hard not to trust, because we have evolved to judge humans—and quickly

continued on p. 95

Digital Object Identifier 10.1109/MSEC.2023.3301262
Date of current version: 10 November 2023

Last Word *continued from p. 96*

ascribe human agency when we see human-like behaviors. Previous computer interfaces were limited, so it was easy to remember that a search window—or a thermostat's interface—wasn't a person. AI agents will interact with the totality of our existence in ways another person would, easily breaking our prior systems of judgment. And so, again, it will need to be trustworthy.

Today's generative AI systems are not trustworthy. We don't know how they are trained. We don't know their secret instructions. We don't know their biases, either accidental or deliberate. All we know is that they are created, at great expense, by corporations that will use every trick they can think of to make them as profitable as possible.

We are going to need to build these generative AI systems and assistive agents with security in mind, from the ground up. This means security from outside attack, of course, but also from the changing business models of the corporations themselves.

Everyone will need their own secure personal data storage. Processing will require secure enclaves. Communications will need to be encrypted. All of this will need to be decoupled, so that no single provider can compromise security. This isn't hard—it's the sort of stuff security engineers do all the time—but we will continuously need to fight against both corporate and government desires for surveillance. This means no backdoors.

Transparency is essential, but only part of the solution. It's possible to poison a model in a way that even someone looking at the training data won't know. And, more generally, the market incentive for the large tech companies to violate our privacy and influence our behavior is just too great. We are going to need to build open-source public models, not licensed from or otherwise controlled by the large tech companies. The development of this transformative technology must not be steered solely by the private sector's near-term financial interests. Generative AI is just too important, too transformative, too relational, and too intimate. We won't truly trust AI unless it's trustworthy, and that requires both secure technology and secure policies. ■

Digital Object Identifier 10.1109/MSEC.2023.3324561