Can you trust AI? Here's why you shouldn't

July 20 2023, by Bruce Schneier and Nathan Sanders

Credit: Pixabay/CC0 Public Domain

If you ask Alexa, Amazon's voice assistant AI system, whether Amazon is a monopoly, it responds by saying it doesn't know. It doesn't take much to make it lambaste the other tech giants, but it's silent about its own corporate parent's misdeeds.

When Alexa responds in this way, it's obvious that it is putting its developer's interests ahead of yours. Usually, though, it's not so obvious

whom an AI system is serving. To avoid being exploited by these systems, people will need to learn to approach AI skeptically. That means deliberately constructing the input you give it and thinking critically about its output.

Newer generations of AI models, with their more sophisticated and less rote responses, are making it harder to tell who benefits when they speak. Internet companies' manipulating what you see to serve their own interests is nothing new. Google's search results and your Facebook feed are filled with paid entries. Facebook, TikTok and others manipulate your feeds to maximize the time you spend on the platform, which means more ad views, over your well-being.

What distinguishes AI systems from these other internet services is how interactive they are, and how these interactions will increasingly become like relationships. It doesn't take much extrapolation from today's technologies to envision AIs that will plan trips for you, negotiate on your behalf or act as therapists and life coaches.

They are likely to be with you 24/7, know you intimately, and be able to anticipate your needs. This kind of conversational interface to the vast network of services and resources on the web is within the capabilities of existing generative AIs like ChatGPT. They are on track to become personalized digital assistants.

As a security expert and data scientist, we believe that people who come to rely on these AIs will have to trust them implicitly to navigate daily life. That means they will need to be sure the AIs aren't secretly working for someone else. Across the internet, devices and services that seem to work for you already secretly work against you. Smart TVs spy on you. Phone apps collect and sell your data. Many apps and websites manipulate you through dark patterns, design elements that deliberately mislead, coerce or deceive website visitors. This is surveillance

capitalism, and AI is shaping up to be part of it.

In the dark

Quite possibly, it could be much worse with AI. For that AI digital assistant to be truly useful, it will have to really know you. Better than your phone knows you. Better than Google search knows you. Better, perhaps, than your close friends, intimate partners and therapist know you.

You have no reason to trust today's leading generative AI tools. Leave aside the hallucinations, the made-up "facts" that GPT and other large language models produce. We expect those will be largely cleaned up as the technology improves over the next few years.

But you don't know how the AIs are configured: how they've been trained, what information they've been given, and what instructions they've been commanded to follow. For example, researchers uncovered the secret rules that govern the Microsoft Bing chatbot's behavior. They're largely benign but can change at any time.

Making money

Many of these AIs are created and trained at enormous expense by some of the largest tech monopolies. They're being offered to people to use free of charge, or at very low cost. These companies will need to monetize them somehow. And, as with the rest of the internet, that somehow is likely to include surveillance and manipulation.

Imagine asking your chatbot to plan your next vacation. Did it choose a particular airline or hotel chain or restaurant because it was the best for you or because its maker got a kickback from the businesses? As with paid results in Google search, newsfeed ads on Facebook and paid

placements on Amazon queries, these paid influences are likely to get more surreptitious over time.

If you're asking your chatbot for political information, are the results skewed by the politics of the corporation that owns the chatbot? Or the candidate who paid it the most money? Or even the views of the demographic of the people whose data was used in training the model? Is your AI agent secretly a double agent? Right now, there is no way to know.

Trustworthy by law

We believe that people should expect more from the technology and that tech companies and AIs can become more trustworthy. The European Union's proposed AI Act takes some important steps, requiring transparency about the data used to train AI models, mitigation for potential bias, disclosure of foreseeable risks and reporting on industry standard tests.

Most existing AIs fail to comply with this emerging European mandate, and, despite recent prodding from Senate Majority Leader Chuck Schumer, the U.S. is far behind on such regulation.

The AIs of the future should be trustworthy. Unless and until the government delivers robust consumer protections for AI products, people will be on their own to guess at the potential risks and biases of AI, and to mitigate their worst effects on people's experiences with them.

So when you get a travel recommendation or political information from an AI tool, approach it with the same skeptical eye you would a billboard ad or a campaign volunteer. For all its technological wizardry, the AI tool may be little more than the same.