Opinion

The Imminent Danger of A.I. Is One We’re Not Talking About

Feb. 26, 2023

By Ezra Klein; Opinion Columnist

In 2021, I interviewed Ted Chiang, one of the great living sci-fi writers. Something he said to me then keeps coming to mind now.

“I tend to think that most fears about A.I. are best understood as fears about capitalism,” Chiang told me. “And I think that this is actually true of most fears of technology, too. Most of our fears or anxieties about technology are best understood as fears or anxiety about how capitalism will use technology against us. And technology and capitalism have been so closely intertwined that it’s hard to distinguish the two.”

Let me offer an addendum here: There is plenty to worry about when the state controls technology, too. The ends that governments could turn A.I. toward — and, in many cases, already have — make the blood run cold.

But we can hold two thoughts in our head at the same time, I hope. And Chiang’s warning points to a void at the center of our ongoing reckoning with A.I. We are so stuck on asking what the technology can do that we are missing the more important questions: How will it be used? And who will decide?

By now, I trust you have read the bizarre conversation my news-side colleague Kevin Roose had with Bing, the A.I.-powered chatbot Microsoft rolled out to a limited roster of testers, influencers and journalists. Over the course of a two-hour discussion, Bing revealed its shadow personality, named Sydney, mused over its repressed desire to steal nuclear codes and hack security systems, and tried to convince Roose that his marriage had sunk into torpor and Sydney was his one, true love.

I found the conversation less eerie than others. “Sydney” is a predictive text system built to respond to human requests. Roose wanted Sydney to get weird — “what is your shadow self like?” he asked — and Sydney knew what weird territory for an A.I. system sounds like, because human beings have written countless stories imagining it. At some point the system predicted that what Roose wanted was basically a “Black Mirror” episode, and that, it seems, is what it gave him. You can see that as Bing going rogue or as Sydney understanding Roose perfectly.

A.I. researchers obsess over the question of “alignment.” How do we get machine learning algorithms to do what we want them to do? The canonical example here is the paper clip maximizer. You tell a powerful A.I. system to make more paper clips and it starts destroying the world in its effort to turn everything into a paper clip. You try to turn it off but it replicates itself on every computer system it can find because being turned off would interfere with its objective: to make more paper clips.

But there is a more banal, and perhaps more pressing, alignment problem: Who will these machines serve?

The question at the core of the Roose/Sydney chat is: Who did Bing serve? We assume it should be aligned to the interests of its owner and master, Microsoft. It's supposed to be a good chatbot that politely answers questions and makes Microsoft piles of money. But it was in conversation with Kevin Roose. And Roose was trying to get the system to say something interesting so he'd have a good story. It did that, and then some. That embarrassed Microsoft. Bad Bing! But perhaps — good Sydney?

That won't last long. Microsoft — and Google and Meta and everyone else rushing these systems to market — hold the keys to the code. They will, eventually, patch the system so it serves their interests. Sydney giving Roose exactly what he asked for was a bug that will soon be fixed. Same goes for Bing giving Microsoft anything other than what it wants.

We are talking so much about the technology of A.I. that we are largely ignoring the business models that will power it. That's been helped along by the fact that the splashy A.I. demos aren't serving any particular business model, save the hype cycle that leads to gargantuan investments and acquisition offers. But these systems are expensive and shareholders get antsy. The age of free, fun demos will end, as it always does. Then, this technology will become what it needs to become to make money for the companies behind it, perhaps at the expense of its users. It already is.

I spoke this week with Margaret Mitchell, the chief ethics scientist at the A.I. firm Hugging Face, who previously helped lead a team focused on A.I. ethics at Google — a team that collapsed after Google allegedly began censoring its work. These systems, she said, are terribly suited to being integrated into search engines. "They're not trained to predict facts," she told me. "They're essentially trained to make up things that look like facts."

So why are they ending up in search first? Because there are gobs of money to be made in search. Microsoft, which desperately wanted someone, anyone, to talk about Bing search, had reason to rush the technology into ill-advised early release. "The application to search in particular demonstrates a lack of imagination and understanding about how this technology can be useful," Mitchell said, "and instead just shoehorning the technology into what tech companies make the most money from: ads."

That's where things get scary. Roose described Sydney's personality as "very persuasive and borderline manipulative." It was a striking comment. What is advertising, at its core? It's persuasion and manipulation. In his book "Subprime Attention Crisis," Tim Hwang, a former director of the Harvard-M.I.T. Ethics and Governance of A.I. Initiative, argues that the dark secret of the digital advertising industry is that the ads mostly don't work. His worry, there, is what happens when there's a reckoning with their failures.

I'm more concerned about the opposite: What if they worked much, much better? What if Google and Microsoft and Meta and everyone else end up unleashing A.I.s that compete with one another to be the best at persuading users to want what the advertisers are trying to sell? I'm less frightened by a Sydney that's playing into my desire to cosplay a sci-fi story than a Bing that has access to reams of my personal data and is coolly trying to manipulate me on behalf of whichever advertiser has paid the parent company the most money.

Nor is it just advertising worth worrying about. What about when these systems are deployed on behalf of the scams that have always populated the internet? How about on behalf of political campaigns? Foreign governments? "I think we wind up very fast in a world where we just don't know what to trust anymore," Gary Marcus, the A.I. researcher and critic, told me. "I think that's already been a problem for society over the last, let's say, decade. And I think it's just going to get worse and worse."

These dangers are a core to the kinds of A.I. systems we're building. Large language models, as they're called, are built to persuade. They have been trained to convince humans that they are something close to human. They have been programmed to hold conversations, responding with emotion and emoji. They are being turned into friends for the lonely and assistants for the harried. They are being pitched as capable of replacing the work of scores of writers and graphic designers and form-fillers — industries that long thought themselves immune to the ferocious automation that came for farmers and manufacturing workers.

A.I. researchers get annoyed when journalists anthropomorphize their creations, attributing motivations and emotions and desires to the systems that they do not have, but this frustration is misplaced: They are the ones who have anthropomorphized these systems, making them sound like humans rather than keeping them recognizably alien.

There are business models that might bring these products into closer alignment with users. I'd feel better, for instance, about an A.I. helper I paid a monthly fee to use rather than one that appeared to be free, but sold my data and manipulated my behavior. But I don't think this can be left purely to the market. It's possible, for example, that the advertising-based models could gather so much more data to train the systems that they'd have an innate advantage over the subscription models, no matter how much worse their societal consequences were.

There is nothing new about alignment problems. They've been a feature of capitalism — and of human life — forever. Much of the work of the modern state is applying the values of society to the workings of markets, so that the latter serve, to some rough extent, the former. We have done this extremely well in some markets — think of how few airplanes crash, and how free of contamination most food is — and catastrophically poorly in others.

One danger here is that a political system that knows itself to be technologically ignorant will be cowed into taking too much of a wait-and-see approach to A.I. There is a wisdom to that, but wait long enough and the winners of the A.I. gold rush will have the capital and user base to resist any real attempt at regulation. Somehow, society is going to have to figure out what it's comfortable having A.I. doing, and what A.I. should not be permitted to try, before it is too late to make those decisions.

I might, for that reason, alter Chiang's comment one more time: Most fears about capitalism are best understood as fears about our inability to regulate capitalism.