AI Is Learning to Manipulate Us, and We Don’t Know Exactly How

Avery Hurt

It’s no secret that the tech giants gather (and buy and sell) tremendous amounts of data about their customers, which is almost all of us. We may rightly worry about how much of our personal data is in the hands of private companies. But we might spend less time thinking about what exactly they do with that data — including using artificial intelligence (AI) to exploit human decision-making.

A Black Box

Humans are pretty good at manipulating each other; in fact, we've likely been engaging in "tactical deception" for thousands of years. But thanks to the assistance of AI, software systems that learn for themselves, humans may be more vulnerable to that coercion ever.

When deployed the right way, artificial intelligence can persuade you to buy something, share a post, vote for a candidate, or do any number of things. Recently, a team of researchers from the Commonwealth Scientific and Industrial Research Organisation, Australia’s federal scientific and research agency, conducted a series of experiments that explored how AI influences human decision-making. The results showed that AI could locate and exploit weaknesses in human decision-making to guide people toward certain decisions. “The implications of this research are potentially quite staggering,” Amir Dezfouli, an expert in machine learning at CSIRO and lead researcher on the study, said in a press release.

Much the same way that a good salesperson (or charming huckster) might get you to do something you might not have otherwise done, these algorithms can get you to click, buy, or share; not only because they

know so much about you, but also because they know what techniques are likely to get you to make one decision rather than another.

And the scary part is that we don't completely understand how AI does it. "The tricky part is that AI is in some ways still a bit of a black box," says Shane Saunderson, who researches human-robot interaction at the University of Toronto. "It's not an explicit machine that says two plus two equals four. It's a machine that you show a bunch of data to, and it analyzes that data for patterns or classifications or insights that it can glean from it. And we don't always know exactly how it's doing that." For example, AI quickly figured out, by collecting and analyzing immense amounts of data, that social media is far more engaging when it plays on negative emotions, and that people react more and engage more with negative content. In recent years, that's had enormous unforeseen consequences.

"This is definitely scary stuff," Saunderson says.

Saunderson describes this as a classic example of the "banality of evil." "There's no nefarious actor that's truly trying to do wrong," he says. "No one at Facebook went out and said, 'Yeah, we want to cause a genocide in Myanmar,' or 'We want to influence the elections at a massive scale.' That was never somebody's intent." The intent, of course, was to sell you stuff — or in the case of Facebook, to keep you engaged on the site so that the companies that buy advertising space can sell you stuff. But the consequences can go far beyond commerce.

For Good or Ill

Dezfouli points out that whether these technologies are used for good or ill depends on how responsibly we design and deploy them. In an attempt to ensure good outcomes, CSIRO and the Australian government developed an ethics framework for AI in government and industry. These (voluntary) principles include much of what you might expect, like "AI systems should respect and uphold privacy rights and data protection." Another tenant says that transparency and responsible disclosure are crucial, so that people can understand when their choices

are being guided and find out when an AI system is engaging with them.

That last one is key, according to Saunderson, who says making AI ethical boils down to transparency. He says that when you interact with a robot or piece of AI you should know, at a minimum, the answers to the following questions:

1) Who owns it or is the interested party behind it?

2) What are its objectives? For example, is it trying to sell you something or convince you that you should take your medicine?

3) What tactics is it using to reach those objectives?

4) What data does it have available?

Sadly, the answers to many of those questions are, for most of us, still a black box.