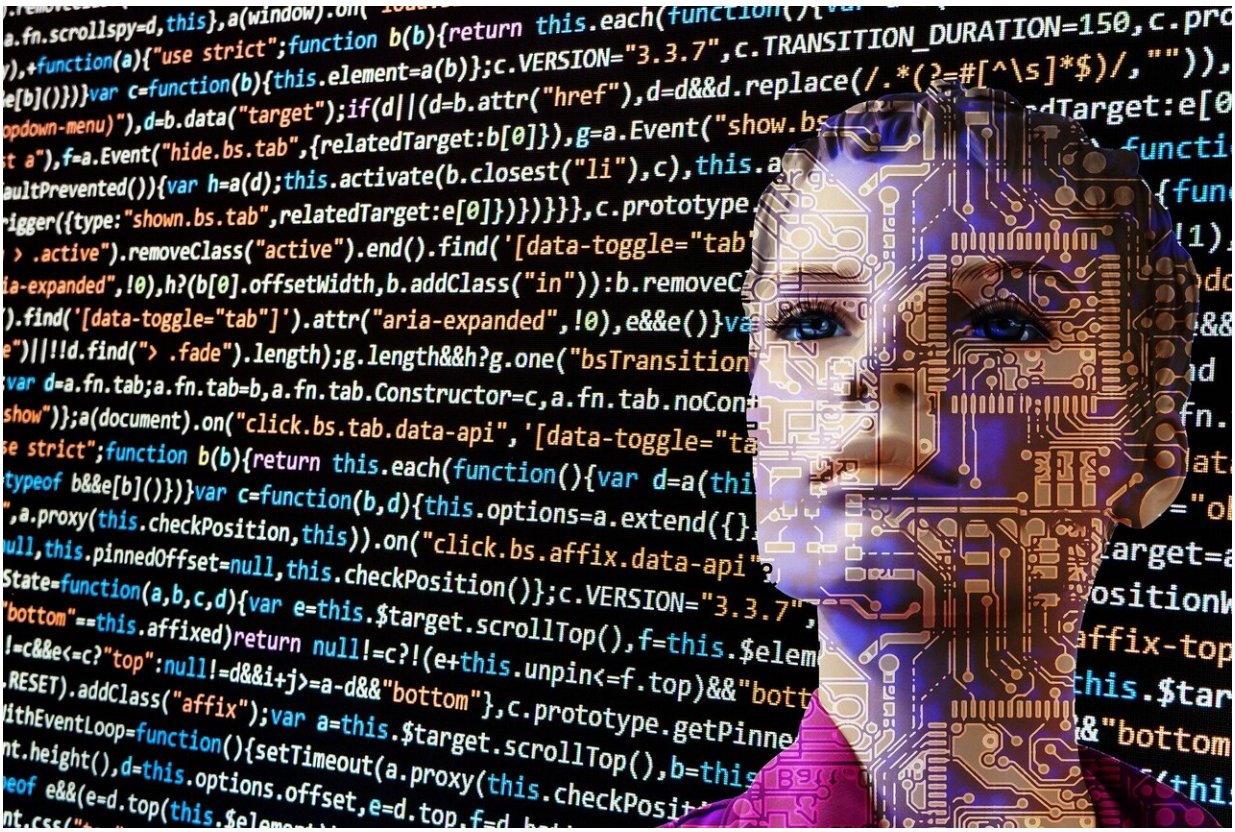
Opinion: AI is not a magic wand—it has built-in problems that are difficult to fix and can be dangerous

June 17 2024, by Niusha Shafiabady



Credit: CC0 Public Domain

By now, all of us have heard and read a lot about artificial intelligence (AI). You've likely used some of the countless AI tools that are

becoming available. For some, AI feels like a magic wand that predicts the future.

But AI is not perfect. A supermarket meal planner in Aotearoa New Zealand gave customers poisonous recipes, a New York City chatbot advised people to break the law, and Google's AI Overview is telling people to eat rocks.

At its core, an AI tool is a particular system that addresses a particular problem. With any AI system, we should match our expectations to its abilities—and many of those come down to how the AI was built.

Let's explore some inherent shortcomings of AI systems.

Trouble in the real world

One of the inherent issues of all AI systems is that they are not 100% accurate in real-world settings. For example, a predictive AI system will be trained using data points from the past.

If the AI then comes across something new—not similar to anything in the training data—it most likely won't be able to make the correct decision.

As a hypothetical example, let's take a military plane equipped with an AI-powered autopilot system. This system will function thanks to its training "knowledge base." But an AI really isn't a magic wand, it's just mathematical computations. An adversary could create obstacles the plane AI cannot "see" because they are not in the training data, leading to potentially catastrophic consequences.

Unfortunately, there is not much we can do about this problem, apart from trying to train the AI for all possible circumstances that we know

of. This can sometimes be an insurmountable task.

Bias in the training data

You may have heard about AI making biased decisions. Usually, bias happens when we have unbalanced data. In simple terms, this means that when training the AI system, we are showing it too many examples of one type of outcome and very few of another type.

Let's take the example of an AI system trained to predict the likelihood a given individual will commit a crime. If the crime data used for training the system mostly contains people from group A (say, a particular ethnicity) and very few from group B, the system won't learn about both groups equally.

As a result, its predictions for group A will make it seem these people are more likely to commit crimes compared to people from group B. If the system is used uncritically, the presence of this bias can have severe ethical consequences.

Thankfully, developers can address this issue by "balancing" the data set. This can involve different approaches, including the use of synthetic data—computer-generated, pre-labeled data built for testing and training AI that has checks built into it to protect against bias.

Being out of date

Another issue with AI can arise when it's been trained "offline" and isn't up to date with the dynamics of the problem it is meant to work on.

A simple example would be an AI system developed to predict daily temperature in a city. Its training data contains all the past information

on temperature data for this location.

After the AI has finished training and is deployed, let's say a severe climactic event disrupts the usual weather dynamics. Since the AI system making the predictions was trained on data that didn't include this disruption, its predictions will become increasingly inaccurate.

The way to solve this issue is training the AI "online," meaning it is regularly shown the latest temperature data while being used to predict temperatures.

This sounds like a great solution, but there are a few risks associated with online training. We can leave the AI system to train itself using the latest data, but it may get out of control.

Fundamentally, this can happen because of chaos theory, which in simple terms means that most AI systems are sensitive to initial conditions. When we don't know what data the system will come across, we can't know how to tune the initial conditions to control potential instabilities in the future.

When the data isn't right

Sometimes, the training data just isn't fit for purpose. For example, it may not have the qualities the AI system needs to perform whatever task we are training it to do.

To use an extremely simplified example, imagine an AI tool for identifying "tall" and "short" people. In the training data, should a person who is 170cm be labeled tall or short? If tall, what will the system return when it comes across someone who is 169.5cm? (Perhaps the best solution would be to add a "medium" label.)

The above may seem trivial, but issues with data labeling or poor data sets can have problematic consequences if the AI system is involved in medical diagnosis, for example.

Fixing this problem is not easy, since identifying the relevant pieces of information requires a great deal of knowledge and experience. Bringing on board a subject matter expert in the data collection process can be a great solution, as it can guide the developers on what types of data to even include to begin with.

As (future) users of AI and technology, it is important for all of us to be aware of these issues to broaden our perspective on AI and its prediction outcomes concerning different aspects of our lives.

Provided by The Conversation

Citation: Opinion: AI is not a magic wand—it has built-in problems that are difficult to fix and can be dangerous (2024, June 17) retrieved 29 June 2024 from https://techxplore.com/news/2024-06-opinion-ai-magic-wand-built.html