Philosophers have studied 'counterfactuals' for decades. Will they help us unlock the mysteries of AI?

January 27 2023, by Sam Baron

Counterfactuals are claims about what would happen were something to occur in a different way. For instance, we can ask what the world would be like had the internet never been developed. Credit: Shutterstock

Artificial intelligence is increasingly being rolled out all around the world to help make decisions in our lives, whether it's loan decisions by banks, medical diagnoses, or US law enforcement predicting a criminal's likelihood of re-offending.

Yet many AI systems are black boxes: no one understands how they work. This has led to a demand for "explainable AI," so we can understand *why* an AI model yielded a specific output, and what biases may have played a role.

Explainable AI is a growing branch of AI research. But what's perhaps less well known is the role philosophy plays in its development.

Specifically, one idea called "counterfactual explanation" is often put forth as a solution to the black box problems. But once you understand the philosophy behind it, you can start to understand why it falls short.

Why explanations matter

When AI is used to make life-changing decisions, the people impacted deserve an explanation of how that decision was reached. This was recently recognized through the European Union's General Data Protection Regulation, which supports an individual's right to explanation.

The need for explanation was also highlighted in the Robodebt case in Australia, where an algorithm was used to predict debt levels for individuals receiving social security. The system made many mistakes, placing people into debt who shouldn't have been.

It was only once the algorithm was fully explained that the mistake was identified—but by then the damage had been done. The outcome was so damaging it led to a royal commission being established in August 2022.

In the Robodebt case, the algorithm in question was fairly straightforward and could be explained. We should not expect this to always be the case going forward. Current AI models using machine-learning to process data are much more sophisticated.

The big, glaring black box

Suppose a person named Sara applies for a loan. The bank asks her to provide information including her marital status, debt level, income, savings, home address and age.

The bank then feeds this information into an AI system, which returns a credit score. The score is low and is used to disqualify Sara for the loan, but neither Sara nor the bank employees know why the system scored Sara so low.

Unlike with Robodebt, the algorithm being used here may be extremely complicated and not easily explained. There is therefore no straightforward way to know whether it has made a mistake, and Sara has no way to get the information she needs to argue against the decision.

This scenario isn't entirely hypothetical: loan decisions are likely to be outsourced to algorithms in the US, and there's a real risk they will encode bias. To mitigate risk, we must try to explain how they work.

The counterfactual approach

Broadly speaking, there are two types of approaches to explainable AI. One involves cracking open a system and studying its internal components to discern how it works. But this usually isn't possible due to the sheer complexity of many AI systems.

The other approach is to leave the system unopened, and instead study its inputs and outputs, looking for patterns. The "counterfactual" method falls under this approach.

Counterfactuals are claims about what would happen if things had

played out differently. In an AI context, this means considering how the output from an AI system might be different if it receives different inputs. We can then supposedly use this to explain why the system produced the result it did.

Suppose the bank feeds its AI system different (manipulated) information about Sara. From this, the bank works out the smallest change Sara would need to get a positive outcome would be to increase her income.

The bank can then apparently use this as an explanation: Sara's loan was denied because her income was too low. Had her income been higher, she would have been granted a loan.

Such counterfactual explanations are being seriously considered as a way of satisfying the demand for explainable AI, including in cases of loan applications and using AI to make scientific discoveries.

However, as researchers have argued, the counterfactual approach is inadequate.

Correlation and explanation

When we consider changes to the inputs of an AI system and how they translate into outputs, we manage to gather information about correlations. But, as the old adage goes, correlation is not causation.

The reason that's a problem is because work in philosophy suggests causation is tightly connected to explanation. To explain why an event occurred, we need to know what caused it.

On this basis, it may be a mistake for the bank to tell Sara her loan was denied because her income was too low. All it can really say with

confidence is that income and credit score are correlated—and Sara is still left without an explanation for her poor result.

What's needed is a way to turn information about counterfactuals and correlations into explanatory information.

The future of explainable AI

With time we can expect AI to be used more for hiring decisions, visa applications, promotions and state and federal funding decisions, among other things.

A lack of explanation for these decisions threatens to substantially increase the injustice people will experience. After all, without explanations we can't correct mistakes made when using AI. Fortunately, philosophy can help.

Explanation has been a central topic of philosophical study over the last century. Philosophers have designed a range of methods for extracting explanatory information from a sea of correlations, and have developed sophisticated theories about how explanation works.

A great deal of this work has focused on the relationship between counterfactuals and explanation. I've developed work on this myself. By drawing on philosophical insights, we may be able to develop better approaches to explainable AI.

At present, however, there's not enough overlap between philosophy and computer science on this topic. If we want to tackle injustice head-on, we'll need a more integrated approach that combines work in these fields.

Commons license. Read the original article.

Provided by The Conversation