Explanations of artificial intelligence: Author proposes model that highlights evidence of fairness

February 15 2023

Credit: Pixabay/CC0 Public Domain

Artificial intelligence (AI) is used in a variety of ways, such as building new kinds of credit scores that go beyond the traditional FICO score. However, while these tools can powerfully and accurately predict

outcomes, their internal operations are often difficult to explain and interpret. As a result, there is a growing demand in ethics and regulation for what is called explainable AI (xAI), especially in high-stakes domains.

In a new article, a professor at Carnegie Mellon University (CMU) suggests that explanations of AI are valuable to those affected by a model's decisions if they can provide evidence that a past adverse decision was unfair. The article is published in *Frontiers in Psychology* for a special issue on AI in Business.

"Recently, legislators in the United States and the European Union have tried to pass laws regulating automated systems, including explainability," says Derek Leben, Associate Teaching Professor of Ethics at CMU's Tepper School of Business, who authored the article. "There are several existing laws that impose legal requirements for explainability, especially with respect to credit and lending, but they are often difficult to interpret when it comes to AI."

In response to demands for explainability, researchers have produced a large set of xAI methods in a short period of time. These methods differ in the type of explanations they can generate, so Leben says we must now ask: What type of explanations are important for an xAI method to produce?

In the article, Leben identifies three types of explanations. One type explains a decision by providing the relative importance of its causal features (for example, "Your income of $40K was the most significant factor in your rejection"). Another type explains a decision by offering a counterfactual change in past states that would have led to a better outcome (for example, "If your salary had been higher than $50K—all else being equal—you would have been approved"). The third type provides practical recommendations on what individuals can do to

improve their future outcomes (for example, "The best way for you to improve your score is to increase your savings by $5K").

While there has been much debate about what type of explanation is most important, Leben supports xAI methods that provide information about counterfactual changes to past states based on what he calls the evidence of fairness view. In this view, individuals affected by a model's decisions (model patients) can and should care about explainability as a means to an end, with the end verifying that a past decision treated them fairly.

Counterfactual explanations can provide people with evidence that a past decision was fair in two ways. The first is to demonstrate that a model would have produced a beneficial decision under alternative conditions that are under the model patient's control (which the author calls positive evidence of fairness). The second is to show that a model would not have produced a beneficial decision when irrelevant behavioral or group attributes are altered (which Leben terms negative evidence of fairness).

Put another way, Leben suggests that xAI methods should be capable of demonstrating that a decision was counterfactually dependent on features that were under the applicant's control (e.g., late payments) and not counterfactually dependent on features that are discriminatory (e.g., race and gender).

Leben says his work has practical implications. Not only can these ideas inform legislative efforts and industry norms around explainability, but they can also be used in other domains. For example, engineers designing AI models and their associated xAI methods can use the evidence of fairness view to help evaluate them.

More information: Derek Leben, Explainable AI as evidence of fair decisions, *Frontiers in Psychology* (2023). DOI:

10.3389/fpsyg.2023.1069426