AI's Got Some Explaining to Do

By **Ariella Brown**
Published: November 14, 2018

KEY TAKEAWAYS
In order to trust the output of an AI system, it is essential to be able to understand its processes and know how it arrived at its conclusions. Explainable AI is the key to eliminating any potential bias.

Source: Sdecoret/Dreamstime.com

Can you trust AI? Should you accept its findings as objectively valid without question? The problem is that even questioning the AI itself would not yield clear answers.

AI systems have generally operated like a black box: Data is input, and data is output, but the processes that transform that data are a mystery. That creates a twofold problem. One is that it is unclear which algorithms' performance are most reliable. The other is that the seemingly objective results can be skewed by the values and biases of the humans who program the systems. This is why there is a need for transparency for the virtual thought processes such systems use, or "explainable AI."

The ethical imperative has become a legal one for anyone subject to GDPR, which impacts not just businesses based in the EU but any that have dealings with people or organizations there. It contains a number of provisions on data protection that extend to EU citizens "the right not to be subject solely to automated decision-making, except in certain situations" and "the right to be provided with meaningful information about the logic involved in the decision."

In other words, it's no longer enough to say, "The algorithm rejected your application." There is a legal mandate to explain the line of thinking that led to the conclusion that has an impact on people's lives. (For more on the pros and cons of AI, check out The Promises and Pitfalls of Machine Learning.)

Biased Results

One concern that some people have raised about algorithmic decisions is that even while standing for objective reasoning, they can reinforce biases. That's the crux of the argument Cathy O'Neil makes in "Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy." The very appearance of objectivity that is associated with big data is what makes it so harmful in its applications that actually reinforce biases.

What she calls "math destruction" is the "result of models that reinforce barriers, keeping particular demographic populations disadvantaged by identifying them as less worthy of credit, education, job opportunities, parole, etc."

She is not alone in finding algorithmic biases. In 2016, Pro Publica shared its findings that algorithms predicted greater recidivism rates for blacks than whites, a factor that translated into different prison sentences for the same types of crimes. A 2017 Guardian article extended the bias to gender as well.

The problem is that these systems get programmed in with far-reaching consequences. In a phone interview Stijn Christiaens, the co-founder and CTO of Collibra, he explained that AI enables "automated decision-making," which can exceed more than 10 thousand decisions per second.

That means that a system set on bad decisions will be making a lot more of them a lot more rapidly than any human could. If the system has a bias, that huge number of decisions can be "damaging to certain populations," with very serious and widespread consequences, Christiaens said.

Care and Feeding of Algorithms

Certainly, there are errors that result from incomplete or poor data. That was the reason some experts quoted in the Guardian article referenced above gave for the biased algorithm results. Oxford University's Sandra Wachter summed it up as follows: "The world is biased, the historical data is biased, hence it is not surprising that we receive biased results."

Along the same lines, Christiaens said, "As it is based on real world observations," AI "observes our biases, and produces sexist or racist outputs." Applying his own terms to what is popularly known as garbage in, garbage out (GIGO), he said the problem could be "the food" that makes up the training data because it is wrong, incomplete or biased itself.

Racist and sexist outcomes can be trained into the system from data that does not adequately represent differences in the population. He offered the case of drawing on training data based on speakers at conferences in which women may have only 20 percent representation. When trained on such skewed representation, the algorithm will have a built-in bias.

AI Alchemy

The AI bias problem is not always due to the data feed, but also the way it works out its decisions. The mystery of those operations so struck Ali Rahimi and Ben Recht that they compared it to alchemy.

While alchemy may have its place, it's not what people want as an answer to their questions about automated decisions with serious consequences. As Rahimi and Recht put it: "But we're now building systems that govern health care and our participation in civil debate. I would like to live in a world whose systems are [built] on rigorous, reliable, verifiable knowledge, and not on alchemy." (For more on AI in health care, see The 5 Most Amazing AI Advances in Health Care.)

Beyond the Black Box: Discovering what Determines the Decisions

This is why some are pushing for a way to introduce transparency into the thinking process of AI systems, having it explain why it arrived at the conclusions that it did. There have been efforts from various places.

A group of three professors and researchers at American universities worked at a solution in 2016 that they called Local Interpretable Model-Agnostic Explanations (LIME). They explain their approach in this video:

Though it was a step in the right direction, the solution didn't work perfectly. And so the research continues, and in light of GDPR, those connected to the EU have a particular interest in achieving explainable AI.

The Artificial Intelligence Lab at the University of Brussel, an institution out of which Christiaens' company emerged, is one of the places devoted to such research. The lab has found ways to work with image recognition and have "the network linguistically explain what is has seen and why" it comes to the conclusions it does about what is in the picture, he said.

"Algorithms always work in the same way," Christiaens explained. "The input data gets translated into features." At the AI lab, they have the means "to drill down and see what occurred in the decision tree." On that basis, it is possible to "see the paths that were followed" to see where something went wrong and then "adjust and retrain."

IBM also directed its attention to the black box problem, and it recently announced the offering of a software service, that will pick up on bias and account for the AI's decisions even while the system is running through the IBM cloud. In addition to the timely alert, it offers suggestions of what data is needed to counteract the biased results.

In addition to the cloud service, IBM is offering consultation for companies who are building machine learning systems to try to reduce biased results in the future. Perhaps other AI experts will also get involved in consulting to help build better systems and offer a check for possible biases that get programmed in.

We have to remember that AI systems are as subject to error as the humans who set them up, and so no one is above giving an account of decisions.

Related Terms

Artificial Intelligence

Explainable Artificial Intelligence