THE WALL STREET JOURNAL.

IDEAS | ESSAY

AI Can't Reason Why

The current data-crunching approach to machine learning misses an essential element of human intelligence

By Judea Pearl and Dana Mackenzie

May 18, 2018 11:30 a.m. ET

Computer programs have reached a bewildering point in their long and unsteady journey toward artificial intelligence. They outperform people at tasks we once felt to be uniquely human, such as playing poker or recognizing faces in a crowd. Meanwhile, self-driving cars using similar technology run into pedestrians and posts and we wonder whether they can ever be trustworthy.

Amid these rapid developments and nagging setbacks, one essential building block of human intelligence has eluded machines for decades: Understanding cause and effect.

Put simply, today's machine-learning programs can't tell whether a crowing rooster makes the sun rise, or the other way around. Whatever volumes of data a machine analyzes, it cannot understand what a human gets intuitively. From the time we are infants, we organize our experiences into causes and effects. The questions "Why did this happen?" and "What if I had acted differently?" are at the core of the cognitive advances that made us human, and so far are missing from machines.

Suppose, for example, that a drugstore decides to entrust its pricing to a machine learning program that we'll call Charlie. The program reviews the store's records and sees that past variations of the price of toothpaste haven't correlated with changes in sales volume. So Charlie recommends raising the price to generate more revenue. A month later, the sales of toothpaste have dropped—along with dental floss, cookies and other items. Where did Charlie go wrong?

Charlie didn't understand that the previous (human) manager varied prices only when the competition did. When Charlie unilaterally raised the price, dentally price-conscious customers took their business elsewhere. The example shows that historical data alone tells us nothing about causes—and that the direction of causation is crucial.

Machine-learning systems have made astounding progress at analyzing data patterns, but that is the low-hanging fruit of artificial intelligence. To reach the higher fruit, AI needs a ladder, which we call the Ladder of Causation. Its rungs represent three levels of reasoning.

The first rung is Association, the level for current machines and many animals; on that rung, Pavlov's dogs learned to associate a bell with food. The next is Intervention: What will happen if I ring a bell, or raise the price of toothpaste? Intervention is different from observation; raising the price unilaterally is different from seeing what happened in the past. The top rung is Counterfactual, which means the ability to imagine results, reflect on one's actions and assess other scenarios. This is the rung that machines need to reach to evaluate and communicate about responsibility, credit, blame and self-improvement. Imagine giving a self-driving car this ability. After an accident, its CPU would ask itself questions like: What would have happened if I had not honked at the drunken pedestrian?

To reach the higher rungs, in place of ever-more data, machines need a model of the underlying causal factors—essentially, a mathematics of cause and effect. A simple element might be: "Liquor impairs people's judgment, and that makes them move in unexpected ways." We can encode this using what scientists now call a causal diagram, in which arrows represent a series of possible causes: Liquor >> Impaired Judgment >> Erratic Motion. Such diagrams are not just pretty pictures, but form the beginning of an algorithm that enables the car to predict that certain pedestrians will react differently to the honking of its horn. They also give us the possibility of "interrogating" the car to explain its process: Why did you honk your horn?

Current machine learning systems can reach higher rungs only in circumscribed domains where the rules are inviolate, such as playing chess. Outside those domains, they are brittle and mistake-prone. But with causal models, a machine can predict the results of actions that haven't been tried before, reflect on its actions, and transfer its learned skills to new situations.

Causal models grew out of work on AI in the 1980s and have spread through health and social sciences, because they can compute at the higher rungs and often unravel statistical paradoxes. They have now come full circle as machine-learning researchers seek more explainable and responsive systems. For instance, scientists at Google and Facebook are examining causal models to analyze online ads to determine whether they make the diffe rence in a product being bought—a counterfactual question.

This is a beginning. When researchers combine data with causal reasoning, we expect to see a mini-revolution in AI, with systems that can plan actions without having seen such actions before; that apply what they have learned to new situations; and that can explain their actions in the native human language of cause and effect.

—Mr. Pearl is a professor of computer science at UCLA and winner of the 2011 Turing Award for his work on probabilistic and causal reasoning. He and Mr. Mackenzie, a mathematics writer, are co-authors of "The Book of Why: The New Science of Cause and Effect," just published by Basic Books.