Chapter 2 in John Brockman (Ed.), *Possible Minds: 25 Ways of Looking at AI*, Penguin Press, 2019.

TECHNICAL REPORT
R-489
May 2019

THE LIMITATIONS OF OPAQUE LEARNING MACHINES

Judea Pearl

***Judea Pearl** is a professor of computer science and director of the Cognitive Systems Laboratory at UCLA. His most recent book, co-authored with Dana Mackenzie, is* The Book of Why: The New Science of Cause and Effect.

As a former physicist, I was extremely interested in cybernetics. Though it did not utilize the full power of Turing Machines, it was highly transparent, perhaps because it was founded on classical control theory and information theory. We are losing this transparency now, with the deep-learning style of machine learning. It is fundamentally a curve-fitting exercise that adjusts weights in intermediate layers of a long input-output chain.

I find many users who say that it "works well and we don't know why." Once you unleash it on large data, deep learning has its own dynamics, it does its own repair and its own optimization, and it gives you the right results most of the time. But when it doesn't, you don't have a clue about what went wrong and what should be fixed. In particular, you do not know if the fault is in the program, in the method, or because things have changed in the environment. We should be aiming at a different kind of transparency.

Some argue that transparency is not really needed. We don't understand the neural architecture of the human brain, yet it runs well, so we forgive our meager understanding and use human helpers to great advantage. In the same way, they argue, why not unleash deep-learning systems and create intelligence without understanding how they work? I buy this argument to some extent. I personally don't like opacity, so I won't spend my time on deep learning, but I know that it has a place in the makeup of intelligence. I know that non-transparent systems can do marvelous jobs, and our brain is proof of that marvel.

But this argument has its limitation. The reason we can forgive our meager understanding of how human brains work is because our brains work the same way, and that enables us to communicate with other humans, learn from them, instruct them, and motivate them in our own native language. If our robots will all be as opaque as AlphaGo, we won't be able to hold a meaningful conversation with them, and that would be unfortunate. We will need to retrain them whenever we make a slight change in the task or in the operating environment.

So, rather than experimenting with opaque learning machines, I am trying to understand their theoretical limitations and examine how these limitations can be overcome. I do it in the context of causal-reasoning tasks, which govern much of how scientists think about the world and, at the same time, are rich in intuition and toy examples, so we can monitor the progress in our analysis. In this context, we've discovered that some basic barriers exist, and that unless they are breached we won't get a real human kind of intelligence no matter what we do. I believe that charting these barriers may be no less important than banging our heads against them.

Current machine-learning systems operate almost exclusively in a statistical, or model-blind, mode, which is analogous in many ways to fitting a function to a cloud of data points. Such systems cannot reason about "what if ?" questions and, therefore, cannot serve as the basis for Strong AI—that is, artificial intelligence that emulates human-level reasoning and competence. To achieve human-level intelligence, learning machines need the guidance of a blueprint of reality, a model—similar to a road map that guides us in driving through an unfamiliar city.

To be more specific, current learning machines improve their performance by optimizing parameters for a stream of sensory inputs received from the environment. It is a slow process, analogous to the natural-selection process that drives Darwinian evolution. It explains how species like eagles and snakes have developed superb vision systems over millions of years. It cannot explain, however, the super-evolutionary process that enabled humans to build eyeglasses and telescopes over barely a thousand years. What humans had that other species lacked was a mental representation of their environment—representations that they could manipulate at will to imagine alternative hypothetical environments for planning and learning.

Historians of *Homo sapiens* such as Yuval Noah Harari and Steven Mithen are in general agreement that the decisive ingredient that gave our ancestors the ability to achieve global dominion about forty thousand years ago was their ability to create and store a mental representation of their environment, interrogate that representation, distort it by mental acts of imagination, and finally answer the "What if?" kind of questions. Examples are interventional questions ("What if I do such-and-such?") and retrospective or counterfactual questions ("What if I had acted differently?"). No learning machine in operation today can answer such questions. Moreover, most learning machines do not possess a representation from which the answers to such questions can be derived.

With regard to causal reasoning, we find that you can do very little with any form of model-blind curve fitting, or any statistical inference, no matter how sophisticated the fitting process is. We have also found a theoretical framework for organizing such limitations, which forms a hierarchy.

On the first level, you have statistical reasoning, which can tell you only how seeing one event would change your belief about another. For example, what can a symptom tell you about a disease?

Then you have a second level, which entails the first but not vice versa. It deals with actions. "What will happen if we raise prices?" "What if you make me laugh?" That second level of the hierarchy requires information about interventions which is not available in the first. This information can be encoded in a graphical model, which merely tells us which variable responds to another.

The third level of the hierarchy is the counterfactual. This is the language used by scientists. "What if the object were twice as heavy?" "What if I were to do things differently?" "Was it the aspirin that cured my headache, or the nap I took?" Counterfactuals are at the top level in the sense that they cannot be derived even if we could predict the effects of all actions. They need an extra ingredient, in the form of equations, to tell us how variables respond to changes in other variables.

One of the crowning achievements of causal-inference research has been the algorithmization of both interventions and counterfactuals, the top two layers of the hierarchy. In other words, once we encode our scientific knowledge in a model (which may be qualitative), algorithms exist that examine the model and determine if a given query, be it about an intervention or about a counterfactual, can be estimated from the available data—and, if so, how. This capability has transformed dramatically the way scientists are doing science, especially in such data-intensive sciences as sociology and epidemiology, for which causal models have become a second language. These disciplines view their linguistic transformation as the Causal Revolution. As Harvard social scientist Gary King puts it, "More has been learned about causal inference in the last few decades than the sum total of everything that had been learned about it in all prior recorded history."

As I contemplate the success of machine learning and try to extrapolate it to the future of AI, I ask myself, "Are we aware of the basic limitations that were discovered in the causal-inference arena? Are we prepared to circumvent the theoretical impediments that prevent us from going from one level of the hierarchy to another level?"

I view machine learning as a tool to get us from data to probabilities. But then we still have to make two extra steps to go from probabilities into real understandingnce—two big steps. One is to predict the effect of actions, and the second is counterfactual imagination. We cannot claim to understand reality unless we make the last two steps.

In his insightful book *Foresight and Understanding* (1961), the philosopher Stephen Toulmin identified the transparency-versus-opacity contrast as the key to understanding the ancient rivalry between Greek and Babylonian sciences. According to Toulmin, the Babylonian astronomers were masters of black-box predictions, far surpassing their Greek rivals in accuracy and consistency of celestial observations. Yet Science favored the creative-speculative strategy of the Greek astronomers, which was wild with metaphorical imagery: circular tubes full of fire, small holes through which celestial fire was visible as stars, and hemispherical Earth riding on turtleback. It was this wild modeling strategy, not Babylonian extrapolation, that jolted Eratosthenes (276-194 BC) to perform one of the most creative experiments in the ancient world and calculate the circumference of the Earth. Such an experiment would never have occurred to a Babylonian data-fitter.

Model-blind approaches impose intrinsic limitations on the cognitive tasks that Strong AI can perform. My general conclusion is that human-level AI cannot emerge solely from model-blind learning machines; it requires the symbiotic collaboration of data and models.

Data science is a science only to the extent that it facilitates the interpretation of data—a two-body problem, connecting data to reality. Data alone are hardly a science, no matter how "big" they get and how skillfully they are manipulated. Opaque learning systems may get us to Babylon, but not to Athens.