To Build Truly Intelligent Machines, Teach Them Cause and Effect

Judea Pearl, a pioneering figure in artificial intelligence, argues that AI has been stuck in a decades-long rut. His prescription for progress? Teach machines to understand the question why.

By Kevin Hartnett

Monica Almeida for Quanta Magazine

Artificial intelligence owes a lot of its smarts to Judea Pearl. In the 1980s he led efforts that allowed machines to reason probabilistically. Now he's one of the field's sharpest critics. In his latest book, "The Book of Why: The New Science of Cause and Effect," he argues that artificial intelligence has been handicapped by an incomplete understanding of what intelligence really is.

Three decades ago, a prime challenge in artificial intelligence research was to program machines to associate a potential cause to a set of observable conditions. Pearl figured out how to do that using a scheme called Bayesian networks. Bayesian networks made it practical for machines to say that, given a patient who returned from Africa with a fever and body aches, the most likely explanation was malaria. In 2011 Pearl won the Turing Award, computer science's highest honor, in large part for this work.

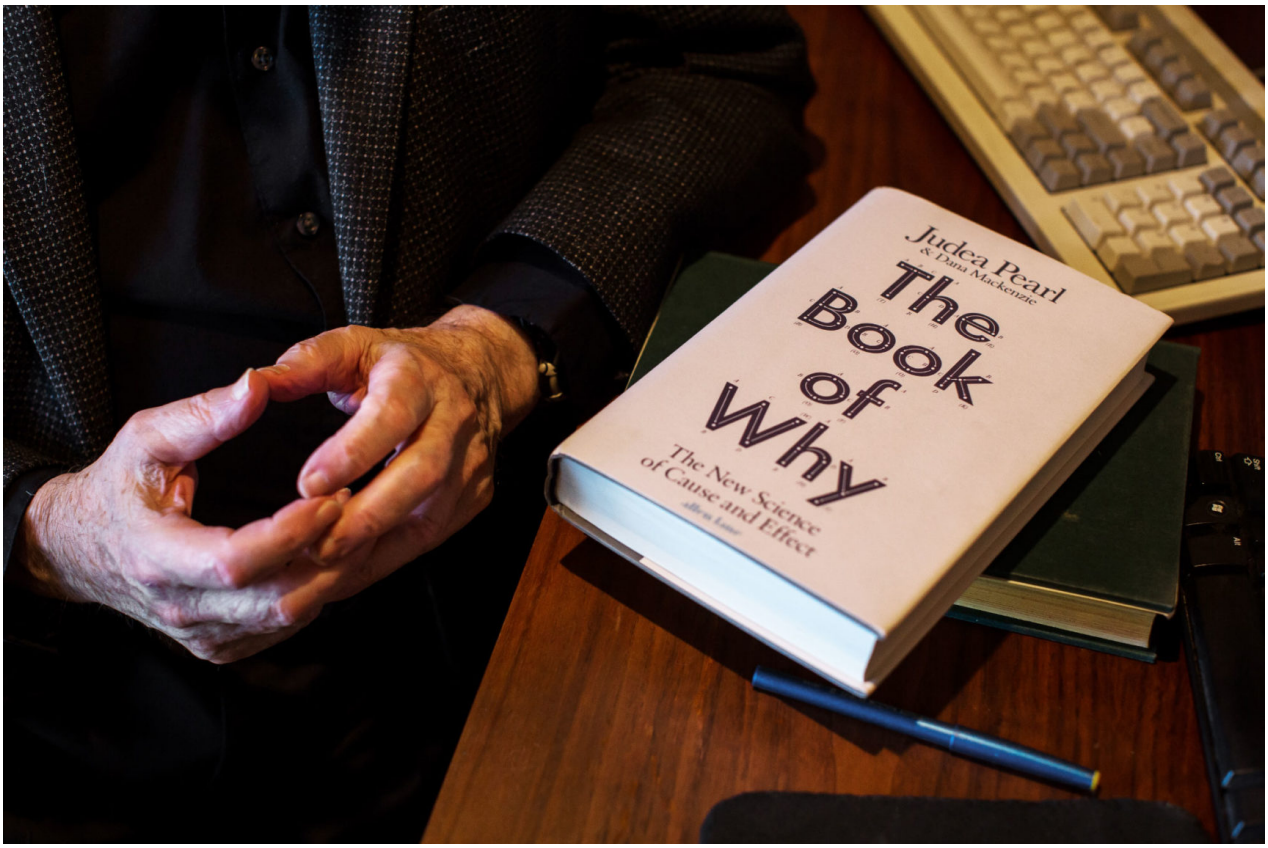
But as Pearl sees it, the field of AI got mired in probabilistic associations. These days, headlines tout the latest breakthroughs in machine learning and neural networks. We read about computers that can master ancient games and drive cars. Pearl is underwhelmed. As he sees it, the state of the art in artificial intelligence today is merely a souped-up version of what machines could already do a generation ago: find hidden regularities in a large set of data. "All the impressive achievements of deep learning amount to just curve fitting," he said recently.

In his new book, Pearl, now 81, elaborates a vision for how truly intelligent machines would think. The key, he argues, is to replace reasoning by association with causal reasoning. Instead of the mere ability to correlate fever and malaria, machines need the capacity to reason that malaria causes fever. Once this kind of causal framework is in place, it becomes possible for machines to ask counterfactual questions — to inquire how the causal relationships would change given some kind of intervention — which Pearl views as the cornerstone of scientific thought. Pearl also proposes a formal language in which to make this kind of thinking possible — a 21st-century version of the Bayesian framework that allowed machines to think probabilistically.

Pearl expects that causal reasoning could provide machines with human-level intelligence. They'd be able to communicate with humans more effectively and even, he explains, achieve status as moral entities with a capacity for free will — and for evil. *Quanta Magazine* sat down with Pearl at a recent conference in San Diego and later held a follow-up interview with him by phone. An edited and condensed version of those conversations follows.

Why is your new book called "The Book of Why"?

It means to be a summary of the work I've been doing the past 25 years about cause and effect, what it means in one's life, its applications, and how we go about coming up with answers to questions that are inherently causal. Oddly, those questions have been abandoned by science. So I'm here to make up for the neglect of science.



Monica Almeida for Quanta Magazine

That's a dramatic thing to say, that science has abandoned cause and effect. Isn't that exactly what all of science is about?

Of course, but you cannot see this noble aspiration in scientific equations. The language of algebra is symmetric: If X tells us about Y, then Y tells us about X. I'm talking about deterministic relationships. There's no way to write in mathematics a simple fact — for example, that the upcoming storm causes the barometer to go down, and not the other way around.

Mathematics has not developed the asymmetric language required to capture our understanding that if X causes Y that does not mean that Y causes X. It sounds like a terrible thing to say against science, I know. If I were to say it to my mother, she'd slap me.

But science is more forgiving: Seeing that we lack a calculus for asymmetrical relations, science encourages us to create one. And this is where mathematics comes in. It turned out to be a great thrill for me to see that a simple calculus of causation solves problems that the greatest statisticians of our time deemed to be ill-defined or unsolvable. And all this with the ease and fun of finding a proof in high-school geometry.

You made your name in AI a few decades ago by teaching machines how to reason probabilistically. Explain what was going on in AI at the time.

The problems that emerged in the early 1980s were of a predictive or diagnostic nature. A doctor looks at a bunch of symptoms from a patient and wants to come up with the probability that the

patient has malaria or some other disease. We wanted automatic systems, expert systems, to be able to replace the professional — whether a doctor, or an explorer for minerals, or some other kind of paid expert. So at that point I came up with the idea of doing it probabilistically.

Unfortunately, standard probability calculations required exponential space and exponential time. I came up with a scheme called Bayesian networks that required polynomial time and was also quite transparent.

Yet in your new book you describe yourself as an apostate in the AI community today. In what sense?

In the sense that as soon as we developed tools that enabled machines to reason with uncertainty, I left the arena to pursue a more challenging task: reasoning with cause and effect. Many of my AI colleagues are still occupied with uncertainty. There are circles of research that continue to work on diagnosis without worrying about the causal aspects of the problem. All they want is to predict well and to diagnose well.

I can give you an example. All the machine-learning work that we see today is conducted in diagnostic mode — say, labeling objects as “cat” or “tiger.” They don’t care about intervention; they just want to recognize an object and to predict how it’s going to evolve in time.

I felt an apostate when I developed powerful tools for prediction and diagnosis knowing already that this is merely the tip of human intelligence. If we want machines to reason about interventions (“What if we ban cigarettes?”) and introspection (“What if I had finished high school?”), we must invoke causal models. Associations are not enough — and this is a mathematical fact, not opinion.

People are excited about the possibilities for AI. You’re not?

As much as I look into what’s being done with deep learning, I see they’re all stuck there on the level of associations. Curve fitting. That sounds like sacrilege, to say that all the impressive achievements of deep learning amount to just fitting a curve to data. From the point of view of the mathematical hierarchy, no matter how skillfully you manipulate the data and what you read into the data when you manipulate it, it’s still a curve-fitting exercise, albeit complex and nontrivial.

Monica Almeida for Quanta Magazine

The way you talk about curve fitting, it sounds like you're not very impressed with machine learning.

No, I'm very impressed, because we did not expect that so many problems could be solved by pure curve fitting. It turns out they can. But I'm asking about the future — what next? Can you have a robot scientist that would plan an experiment and find new answers to pending scientific questions? That's the next step. We also want to conduct some communication with a machine that is meaningful, and meaningful means matching our intuition. If you deprive the robot of your intuition about cause and effect, you're never going to communicate meaningfully. Robots could not say "I should have done better," as you and I do. And we thus lose an important channel of communication.

What are the prospects for having machines that share our intuition about cause and effect?

We have to equip machines with a model of the environment. If a machine does not have a model of reality, you cannot expect the machine to behave intelligently in that reality. The first step, one that will take place in maybe 10 years, is that conceptual models of reality will be programmed by humans.

The next step will be that machines will postulate such models on their own and will verify and refine them based on empirical evidence. That is what happened to science; we started with a geocentric model, with circles and epicycles, and ended up with a heliocentric model with its

ellipses.

Robots, too, will communicate with each other and will translate this hypothetical world, this wild world, of metaphorical models.

When you share these ideas with people working in AI today, how do they react?

AI is currently split. First, there are those who are intoxicated by the success of machine learning and deep learning and neural nets. They don't understand what I'm talking about. They want to continue to fit curves. But when you talk to people who have done any work in AI outside statistical learning, they get it immediately. I have read several papers written in the past two months about the limitations of machine learning.

Are you suggesting there's a trend developing away from machine learning?

Not a trend, but a serious soul-searching effort that involves asking: Where are we going? What's the next step?

That was the last thing I wanted to ask you.

I'm glad you didn't ask me about free will.

In that case, what do you think about free will?

We're going to have robots with free will, absolutely. We have to understand how to program them and what we gain out of it. For some reason, evolution has found this sensation of free will to be computationally desirable.

In what way?

You have the sensation of free will; evolution has equipped us with this sensation. Evidently, it serves some computational function.

Will it be obvious when robots have free will?

I think the first evidence will be if robots start communicating with each other counterfactually, like "You should have done better." If a team of robots playing soccer starts to communicate in this language, then we'll know that they have a sensation of free will. "You should have passed me the ball — I was waiting for you and you didn't!" "You should have" means you could have controlled whatever urges made you do what you did, and you didn't. So the first sign will be communication; the next will be better soccer.

Now that you've brought up free will, I guess I should ask you about the capacity for evil, which we generally think of as being contingent upon an ability to make choices. What is evil?

It's the belief that your greed or grievance supersedes all standard norms of society. For example, a person has something akin to a software module that says "You are hungry, therefore you have permission to act to satisfy your greed or grievance." But you have other software modules that instruct you to follow the standard laws of society. One of them is called compassion. When you

elevate your grievance above those universal norms of society, that’s evil.

So how will we know when AI is capable of committing evil?

When it is obvious for us that there are software components that the robot ignores, consistently ignores. When it appears that the robot follows the advice of some software components and not others, when the robot ignores the advice of other components that are maintaining norms of behavior that have been programmed into them or are expected to be there on the basis of past learning. And the robot stops following them.