The “Book of Why” by Judea Pearl was a great primer to learn about the historical relationship between statistics, causality, and artificial intelligence. It answered many questions which I have asked myself in the past regarding the “hole in math” around causal reasoning. When I was learning statistics and probability theory in school, the only mention of causation was in the phrase *“Correlation is not Causation!”.* I would say that this phrase has become a meme among people who consider themselves intellectuals, and typically use it as a blunt rhetorical device. But the begging question which should follow, *“Then what is causation according to math?!”* has never been answered for me.

As Judea Pearl wrote, statistics seemed to be viewing causality as an artifact of subjective perception. In hindsight, I must say that this view probably translated into my math class in school and contributed to a personal feeling of detachment towards statistics. The second time I learned statistics was in a course on Applied Statistics for Natural Language Processing. That time, it felt much easier to like statistics, because the course allowed us to interpret the math in a cause-effect context. For example, in Language Modeling, statistics can be applied to calculate the most likely continuing word based on a sequence of previous words. Given the sequence “Obama was the President of the United”, “States” would be the most likely continuation. It would not be frowned upon to say that the occurrence of the specific previous words “caused” the likely continuation candidate. I have never hesitated to use causal terminology in the context of Bayes’ rule, but I probably never should have, at least according to Pearson, Fisher, and purists in the history of statistics.

It was unexpectedly relieving for me to read Judea’s introductions of Pathway Analysis, the Inference Engine, Causal Diagrams, and the Do-operator. Finally, the reason why I felt originally disconnected from my first experience with statistics was uncovered, and the hole in math around causality was filled.

However, for me, the most interesting concept which Judea introduced in the “Book of Why” was the Ladder of Causation. The proposed classification of Cognitive Systems into three nested categories of ability – Seeing, Doing, and Imagining – was as profound to me as learning about the Chomsky hierarchy of grammars. I think that this ladder is a great contribution towards the definition of intelligence, moving towards a more refined and differentiated view of what an “Intelligent” agent must be able to do to behave and evolve in a certain way.

Surprisingly for me, Judea classified all current Machine Learning/Deep Learning techniques squarely on the first rung of the ladder – Seeing – asserting that they do not work in cause-effect frameworks of thought, and they do not consider the effects of their interventions. This leads me to discuss the second question: How can it be that machine learning is so popular and successful in many areas, even if the techniques are allegedly oblivious to causal dependencies?

First, I must loudly voice my disagreement with this assertion, and I consider it to be the weakest part of the book. Judea Pearl is conflating Data Science, Data Mining, and Deep Learning. And within Deep Learning, he is conflating all techniques, be it Supervised, Unsupervised or Reinforcement Learning.

I can get behind the idea that a statistical evaluation of correlating variables in a dataset does not lead beyond the first rung of the Ladder of Causation. But in my opinion, this is does not mean that a Deep Learning (DL) model is blind to causes and effects. In the most basic version of DL, a model is asked to learn a function f:X->Y which infers labels Y from features X by adjusting variables which control how the features and the labels are associated. However, this goes far beyond observing correlations: A *well-generalized* Neural Network model is in essence a Causal Diagram for the cause-effect structure between the features and the labels. Of course, usually a network does not generalize well. It will associate the wrong features with the wrong labels and predict the moon to be a traffic light.

But here I already see the first answer to the second question: I think that Neural Networks are successful because they *do* operate on a notion of cause and effect, contrary to the belief of Judea Pearl. This still leaves basic Supervised Feed-Forward Networks on rung 1, because they do not have any way of performing experiments to test their causal assumptions using interventions, and therefore, they cannot generate new evidence (training data) for themselves.

However, there is a machine learning technique which gives models the ability to test their causal hypotheses: Reinforcement Learning. In Reinforcement Learning, an agent which is placed within some world (e.g., Tetris) is asked to perform a task (e.g., winning the game) by executing a sequence of actions (a "policy") which the agent believes to lead to success ("reward"). If the actions do not lead to reward, the agent must adjust its policy. In this case, I think that we are completely operating on rung two of the ladder of causation. A Reinforcement Learning agent is clearly evolving a causal model of its world to predict the relationship between its actions and the reward it is receiving.

So, in my opinion, the real question is about the third rung: Imagining alternate worlds and counterfactual scenarios. This is somewhat in line with the predictions by Jeff Hawkins from "Theory of a Thousand Brains". According to Hawkins, the missing ingredient for AGI are distributed world models based on Reference Frames, inspired by the architecture of cortical columns, which constantly predict incoming stimuli. Lisa Feldmann-Berrett also talked about this "observation before stimulus" concept in great detail in "7½ Lessons about the brain". In such an architecture, the process of "imagining" a counterfactual scenario would simply mean to "unplug" the sensorimotor links from an embodied policy-reward-model! The third rung of the ladder is clearly within reach.

I am not sure how Judea Pearl could ignore this, but it really seems like he has missed most of the progress in Machine Learning research from the last 10 years. I cannot explain it otherwise. The conflation of different machine learning techniques is my biggest issue with his book. I feel like the author is trying to make a point about Deep Learning, but does not really understand what Deep Learning, especially Reinforcement Learning, actually is. I also fear that when we reach AGI, Judea might not be happy with the result. An intelligent cognitive agent might be able to talk in causal terms about their motivations for their actions, but I do not think that this will make their underlying models ("brains") a lot more causally explainable than the inner workings of a human brain. There is probably a big difference between a subjective, conscious *explanation of Why*, and the actual *physical mechanics of Why*.

In my opinion, the part of Deep Learning that is its greatest source of uncertainty about the models it produces is nowhere to be replaced: Gradient Descent. The lack of interpretability in Gradient Descent is the reason why we cannot be sure whether our DL models generalize. So, I fully believe we might reach AGI, and we still we might barely understand more of what an agent is “thinking” than we understand current DL models, or even the human brain. But I really hope to be proven wrong!

7571 characters (counted with MS Word).