TECHNICAL REPORT
R-490
March 2022

GENERAL PERSPECTIVE

Interviewer: Inferring causal effects from data involves many steps. Where do you think your work fits within this overall process?

Pearl: I seek to understand the conditions under which such inference is theoretically possible, allowing of course for partial scientific knowledge to guide the inference. My focus has been on a class of models called "nonparametric" which enjoy two unique features: (1) They capture faithfully the kind of scientific knowledge that is available to empirical researchers and (2) they require no commitment to numerical assumptions of any sort. Leveraging these models, I have focused on the problem of identification, rather than estimation. This calls for transforming the desired causal quantity into an equivalent probabilistic expression (called estimand) that can be estimated from data. Once an estimand is derived, the actual estimation step is no longer causal, and can be accomplished by standard statistical methods. This is indeed where machine learning excels, unlike the identification step in which machine learning and standard statistical methods are almost helpless. It is for this reason that I focus on identification – this is where the novelty of causal thinking lies, and where a new calculus had to be developed.

HISTORICAL PERSPECTIVE

Interviewer: What is your perspective of the history of the causal inference movement, and how the movement came to where it is today?

Pearl: My perspective comes through the lens of a computer scientist. I look at this movement as a struggle to develop a mathematical language for capturing cause effect relationships, so that we can express our assumptions faithfully and transparently, derive their logical implications and combine them with data. It's really a wedding between two non-intersecting languages, one is the language of cause and effect, the other is the language of data, namely statistics (Pearl, 2019a).

The wedding occurred quite late in the history of science because science had not been very kind to causality. It has revolved around the symmetric equality sign '=' of algebra, and thus deprived us of a language to capture the asymmetry of causal relationships. Such a language was developed in the past three decades, using graphs, and it now enables us to answer causal and counterfactual questions with algorithmic precision.

Graphs are new mathematical objects, unfamiliar to most researchers in the statistical sciences, and were of course rejected as "non-scientific ad-hockery" by top leaders in the field (Rubin, 2009). My attempts to introduce causal diagrams to statistics (Pearl, 1995, 2009) have taught me that inertial forces play at least as strong a role in science as they do in politics. That is the reason that non-causal mediation analysis is still practiced in certain circles of social science (Hayes, 2017), "ignorability" assumptions still dominate large islands of research (Imbens & Rubin, 2015), and graphs are still tabooed in the econometric literature (Angrist & Pischke, 2015). While most researchers today acknowledge the merits of graph as a transparent language for articulating scientific information, few appreciate the computational role of graphs as "reasoning engines," namely, bringing to light the logical ramifications of the information used in their construction. Some economists even go to great pains to suppress this computational miracle (Heckman & Pinto, 2015).

Although statistics began in the 1800s, Sewall Wright was the first person (in 1920) to put down mathematically the assumption that X causes Y and not the other way around (Wright S., 1921). Using "path diagrams" he was able to articulate causal assumptions mathematically, communicate them and defend them on scientific grounds. Moreover, given the structure of the diagram and its path coefficients, he could calculate correlations among measured variables. Subsequently he worked backwards and found the coefficients from the correlations.

In retrospect, Wright's exercise was remarkable (Pearl & Mackenzie, 2018). Everybody jumped on him for doing the impossible, extracting causation from correlation. But he responded admirably, claiming that his causal conclusions emerge not from correlations alone, but from a combination of correlations and causal assumptions. This is the same philosophy that rules causal inference today. Yet even today, many researchers still have a hard time understanding how path diagrams can do things that correlations alone cannot (Sobel, 2009); they haven't been taught how to read the causal assumptions that are so vividly displayed in the diagram.

In 1923, and independently of Wright, Jersey Neyman introduced another notation for causal effects, in the context of controlled experiments (Neyman, 1923). His notation invoked counterfactuals (more accurately, potential outcomes): Y_1 is what you would see if you apply treatment 1 and Y_0 is what you would see if you apply treatment 0. The rules of probability can now be applied to those counterfactual entities as if they were ordinary variables.

Fisher, the inventor of randomized experiments, did not use Neyman's notation. He used intuition to claim, not prove, that randomization gives you what you want (Fisher, 1926). Note that farmers could not care less about randomization. They want to know what the yield would be if they applied fertilizer 1 or 0 to the entire field, not to randomly selected lots. But though he did not have the notation to express what the farmer wants, Fisher nevertheless convinced the entire statistical community that, if you randomize, you get (on the average) what the farmer wants (i.e., the treatment effect). His argument was so compelling that statisticians accepted it without a mathematical proof.

Within economics, the use of causal notation started in 1928 with Philip Wright (the father of Sewall) who used structural equations to invent the method of instrumental variables (Wright P., 1928). Haavelmo (1943) later looked at the practice of economic modelling and noticed that the models invoked equations of peculiar character. He was the first to ask "what does this equation say?" Before him, people understood intuitively that the equality sign is not an ordinary equality, and that the equation said something profound and extra-statistical about how the economy works, but no one dared name it "causal effect." They could not articulate what causal assumptions are conveyed by an economic equation. Haavelmo was the first to assert that, when an economist puts down an economic equation, he/she has an experiment in mind (Pearl, 2015a). On the right-hand side, you have controlled variables and on the left-hand side you have a function of those variables. He thus assigned causal meaning to the equation, and proceeded to devise a mathematical procedure for combining several equations and deriving causal effects. He essentially said that one should modify the right-hand side of the equation to obtain the desired causal effect (Pearl, 2015a).

In 1960, Strotz and Wold "wiped out" equations from the model to simulate price fixing (Strotz & Wold, 1960). This was the second step in the transition from algebraic to graphical methods. But it had to wait for Spirtes, Glymour and Scheines in 1991 to give it graphical representation by removing arrows from the diagram (Spirtes, Glymour, & Scheines, 1993).

In the 1960's Blalock and Duncan introduced causal inference to social science (Blalock, 1962, 1964; Duncan, 1966). Like Wright, they used path diagrams to get partial correlations, and then used inversion to get back the path coefficients. The field exploded after this. Every social or behavioral scientist became an expert in partial correlation, path coefficients or path diagrams. Unfortunately, they didn't have *d*-separation. This meant they could not read vanishing partial correlations from the diagram, and could not see all the nice things that those partials imply, like identification, model equivalence and more. Eventually in 1975, Jöreskog invented LISREL (Linear Structural Relations), a software package that fits a model directly to the data and provides you with a degree of fitness (Jöreskog & Sörbom, 1978). Practitioners forgot the causal meaning of the equations and thought they were doing statistics. Some even campaigned against "thinking or using terms such as cause and effect" (Muthen, 1987). That is the way some social scientists still operate (Bollen & Pearl, 2013), although there has been a great revival of causal inference through the book of Morgan and Winship (2007, 2015).

Also, in the 1970s, Rubin noticed that Neyman's notation can be used for causal inference not only in experimental studies, but also in observational studies (Rubin, 1974). He was one of the first (1980) to put down the consistency rule in the form of an equation $Y = XY_1 + (1 - X)\ Y_0$, which he considered to be an "assumption" (Rubin, 1980). This important equation connects the hypothetical counterfactuals Y_1 and Y_0 with an observed quantity, namely Y. Lewis and Stalnaker came out with a theory of counterfactuals 8 years earlier, using possible worlds semantics (Stalnaker, 1980), and Gibbard and Harper proved (1976) that the consistency rule is actually a theorem that follows from Lewis semantics (Gibbard & Harper, 1981).

In the 1980s, Greenland and Robins took the counterfactual notation and used it to classify experimental units into 4 response types: Doomed, Causative, Preventive and Immune (Greenland & Robins, 1986). They defined confounding in terms of those response types, which was the first formal definition of "confounding." Previous attempts to define confounding in terms of statistical vocabulary failed of course because confounding is a causal concept. Unfortunately, assessing the proportion of people in each response type is hard, because individuals are not labeled by type.

In 1986, Robins used Neyman notation to derive his *G*-formula, assuming independencies among counterfactuals. (i.e., ignorability). It provided answers to questions such as, "If you have a collection of temporally ordered variables and you make the assumptions of ignorability at every stage in time, what would be the effect of a sequence of interventions?" He showed that, if you make the assumption that every variable is randomized at every stage, given the past, you can assess the effect of interventions given pre-intervention probabilities (Robins, 1986).

In 1991, Spirtes, Glymour and Scheines derived the same formula using Strotz and Wold "wiping out" operation on graphs (Spirtes, Glymour, & Scheines, 1993). This had the advantage of basing the assumptions on meaningful relationships between observed variables, as opposed to opaque conditional independencies among counterfactuals. The formula was still limited though to models with no unobserved confounders, also known as "Markovian models." Two years later, I presented the back-door criterion which facilitated the identification of causal effects in "Semi-Markovian models," namely, feedback-free models loaded with unobserved confounders (Pearl, 1993). The back-door criterion made use of the *d*-separation condition which was developed a decade earlier for Bayesian networks (Pearl, 1986).

Bayesian networks were developed to perform probabilistic prediction and retrodiction (diagnosis) using graphical models of conditional independencies. We thought about causation, but we did not dare put down causal assumptions explicitly. Instead we wrote down probabilistic relationships between diseases, symptoms and treatments. We believed we were doing statistics, not causation (Pearl, 1988). Strangely, however, the models were always arranged with parents being causes and the child being an effect; they were actually causal diagrams. Thus, all the knowledge acquired for Bayesian networks turned out applicable to causal diagrams as well, which helped immensely in the development of the *do*-calculus and counterfactual analysis (Balke & Pearl, 1994ab; Pearl, 1994).

I once called *d*-separation "A gift of the Gods," because it is the only bridge we have between the causal assumptions in our model and what we can expect to observe in our data. Scientific communities that adopted the *d*-separation (e.g., epidemiology) have flourished, those that did not, have stayed behind. I am sure historians of 21st century science will take note of this connection.

Thrilled by the power of the back-door criterion, I thought that the language of causality deserves a calculus of its own, that is, a set of procedures to answer any causal question from any model. Given an arbitrary graph, I sought a mechanical procedure to get an answer to the question "If I intervene on *X*, what will happen to *Y*?" The *do*-calculus came out at the right time, because it showed us what can be done beyond back-door adjustments (Pearl, 1994). The front-door criterion was one of its first fruits and *The Book of Why* describes the excitement caused by its discovery (Pearl & Mackenzie, 2018). Another fruit was the sequential version of the back-door criterion, which Jamie Robins and myself derived in 1995. It identifies the effects of time varying treatments (Pearl & Robins, 1995), and demystified the conditions under which Robins's *G*-formula is valid.

In 1995, Phil Dawid had the courage to overrule all negative reviewers and publish my paper in *Biometrika* (Pearl, 1995). He thought that the field needs to hear about this new way of dealing with causal effects. Such editorial courage is rare these days.

Starting about the same time, Rubin's potential outcome framework became popular in several segments of the research community, mostly among economists and political scientists. These researchers talked "conditional ignorability" to justify their methods, though they could not tell whether it was true or not. Conditional ignorability gave them a formal notation to state a license to use their favorite estimation procedure even though they could not defend the assumptions behind the license. This practice of relying on a priori licenses continues today. It is hard to believe that something so simple as a graph could replace the opaque concept of "conditional ignorability" that people find agonizing and incomprehensible. The back-door criterion made it possible, which was immediately recognized in epidemiology (Greenland, Pearl, & Robins, 1999), though not in all fields (Heckman & Pinto, 2013; Rubin, 2009).

By 1996, I started writing my book Causality. Many problems have been solved since, and the *do*-calculus is now proven to be complete (Shpitser & Pearl, 2006). This means that if the *do*-calculus tells you that it cannot identify a certain causal effect, there is no other method that can identify it non-parametrically, unless you strengthen or refine the assumptions. To elaborate, the *do*-calculus is a set of rules for manipulating causal expressions, the aim of which is to remove the *do*-operator from such expressions and reduce them to statistical estimands. For example, if a certain pattern holds in the graph, you can exchange observation with an action. Another pattern might allow you to remove an action, or remove a variable all together from your expression. An analogy would be symbolic integration in calculus. With a rich set of transformations such as integration by part and integration by substitution, you can simplify the integrand into sums of integrable functions.

In reality, this is still a difficult problem even if you have armed yourself with many tricks of integration, because you don't know what trick to use at any given point of time. Should you use the trick of integration by part or integration by substitution? If substitution, what function should you substitute for *X*, cosine, tangent, logarithm, or exponential? So, you see, having a calculus does not mean that you have an effective procedure to get the answer. We need more than that. The calculus is good for verifying the answer, not for finding it. If you give me a guess of what the identified causal effect looks like, I can prove it to you immediately using the three rules of *do*-calculus. But to find the sequence of rule applications is a difficult problem. Fortunately, we now have an algorithm that just goes ahead and gets us the answer, and exits with failure whenever the answer does not exist (non-identifiability) (Shpitser & Pearl, 2006).

Causal inference can be classified into two distinct classes of problems: predicting effects of interventions and reasoning about causes of effects. The first is formalized by the *do*-calculus while the second requires counterfactual thinking, that is, predicting what the future would be like had the past been different from what it actually was. I consider the algorithmization of counterfactuals (Balke & Pearl, 1994ab) to be one of the crowning achievements of contemporary work on causality. It means that when I specify how the world works, I don't have to think about counterfactuals, ignorability, conditional ignorability or whether one counterfactual is independent of another given a third. No mortal can cognitively handle those counterfactual properties. Instead, a researcher simply writes down a set of structural equations similar to those used by economists, and all counterfactuals are then computed automatically. Every structural equation model determines the "truth value" of every counterfactual sentence. Accordingly, one can compute whether the counterfactual Y_x is independent of any other variable in the model, and you can condition Y_x not only on pre-treatment covariates but also on post-treatment covariates. In this way we can estimate "causes of effects" not merely "effects of causes" (Pearl, 2015b).

Joe Halpern, a computer scientist at Cornell, made a major contribution in 2008 by constructing a complete set of axioms for structural counterfactuals (Halpern, 2008). Such a set can tell us if two interpretations of counterfactuals are equivalent or not. In particular, Halpern's axiomatization proved that Rubin's potential outcome framework is logically equivalent to that of causal graphs. This means that two investigators who explore the same question, on the same data with the same assumptions, will get the same result whether they use Rubin's model or a structural equations model. A theorem in one is a theorem in the other. An assumption in one is an assumption in the other.

GENERALIZABILITY

Interviewer: Researchers in causal inference are starting to examine external validity, also known as generalizability or transportability. Can you describe your approach using selection diagrams, and what questions they address that causal diagrams do not?

Pearl: Causal diagrams describe one population. If you want to specify disparity between two populations, you need to talk about two diagrams. But in many cases, the structures of the two diagrams are the same and the disparity lies in the strengths of the causal relationships. A selection diagram adds a node to one of the diagrams saying: "Here is a factor that creates disparity." We denote such a node by a square. You add these nodes to the causal diagram to mark where the two populations differ. In other words, if there is a square node into variable *Z*, it means that the two populations may differ in terms of the response of *Z* to its parents.

For example, if the age distribution in Los Angeles is different from that of Hawaii, I would add an arrow into Age. It results is a DAG annotated with squares wherever suspicion exists about the homogeneity of the two populations. We now call for *do*-calculus to tell us what we need to know: "Can we generalize what we learned in one population to the other?" The *do*-calculus will manipulate the expressions pertaining to one population and bring them to a format that answers the question (Pearl & Bareinboim, 2014). Moreover, Elias Barenboim has devised algorithmic methods to generate the answer directly, thus bypassing the *do*-calculus (Bareinboim & Pearl, 2013). In Elias's software, the input is a selection diagram and the output will tell us if we can answer the question or not. If we can, it will tell us what information we must acquire from each study, and how to combine them properly. If we combine them the way we are told, we will obtain a consistent estimate of our target answer in the target population (Bareinboim & Pearl, 2016).

I believe history will confirm my current assessment that, following centuries of round-about speculations and wishful thinking, the problem of external validity (and generalizability) has finally been formalized mathematically and now has a path forward to practical applications.

CAUSAL DIAGRAMS, COUNTERFACTUALS AND POTENTIALS OUTCOMES

Interviewer: You said the approach using DAGs and potential outcomes are similar a couple of times. Can you outline what you believe are the differences?

Pearl: I said they are “logically equivalence,” not “similar.“ An analogy would be solving a geometrical problem in polar vs. Cartesian coordinates. Rubin’s framework, known as “potential outcomes,” differs from the structural account in the language in which problems are specified, and hence, in the clarity of articulating what we know and the mathematical tools available for deriving what we wish to know. In the potential outcome framework, problems are defined algebraically as assumptions about counterfactual independencies, also known as “ignorability assumptions.” These types of assumptions are too complicated to interpret or verify by unaided judgment. In the structural framework, on the other hand, problems are defined in the language in which scientific knowledge is stored – causal graphs. Dependencies of counterfactuals, if truly needed, can be deduced from the graph, but in almost all cases they can be replaced by causal dependencies among observables, which are vividly displayed in the graphs.

The reasons some statisticians and economists still prefer the algebraic potential outcome approach are puzzling to me. But as a student of the history, I attribute them to natural resistance to a new language, conformity to traditional cultures, and allegiance to tightly guarded communities.

The advantages of the structural approach can be summarized along three dimensions: *Transparency*, *Power* and *Testability*.

Transparency stands for the ability of a researcher to a) remember assumptions, b) judge their plausibility, c) determine their consistency and, most importantly, d) determine if a set of articulated assumptions is compatible with the requirements of a given identification strategy. Typical identification strategies are “adjustment for covariates” or “instrumental variables.”

Power measures the space of problem instances for which an identification strategy can be found. For example, DAGs together with *do*-calculus can discover *all* identification strategies applicable to a given (nonparametric) interventional problem. The front-door exemplifies an identification strategy that goes beyond “adjustment for covariates.”

Testability stands for one’s ability to determine if the modeling assumptions are compatible with the available data. In DAGs, we have the *d*-separation criterion, which translate immediately into tests of compatibility with data. In potential outcomes, testability requires non-trivial derivations (Pearl, 2014a).

When compared along these three dimensions, the advantages of the structural framework shine uncontested. Unfortunately, only a handful of researchers have taken the time to compare the solution of simple problems, side by side, in the two frameworks, as they are often invited to do. Instead, the weaknesses of the potential outcomes approach are usually glossed over by assuming conditional ignorability a priori and leaving the identification task to the mercy of chance.

NON-MANIPULABLE VARIABLES AND CAUSATION

Interviewer: Can you clarify your thoughts on considering non-manipulable variables like race as a “cause”?

Pearl: The mantra "No causation without manipulation" (Holland, 1986), represents another hang-up of the potential outcome community, which has not been able to liberate itself from the original setup in which potential outcomes were first defined. That setup required a comparison of hypothetical outcomes of conceptually manipulable "treatments." Things are fundamentally different in the structural framework where potential outcomes are defined by (surgeries over) models of reality. There is nothing to prohibit one from taking a model, delete arrows entering variables such as race or blood-pressure then compute and communicate properties of the modified model, for example, $Q=P(y|do(x))$.

For non-manipulable X, this scheme raises two immediate questions. First: What useful information does Q convey aside from being a mathematical property of our model? Second, assuming that Q conveys an important feature of reality, how can we test it empirically? And if we cannot test it, is it part of science?

In a recent paper (Pearl, 2019b) I address precisely these two questions and show that causal effects defined on non-manipulable variables have empirical semantics along three dimensions. First, they provide important information about causal effects of *manipulable* variables. Second, they may facilitate the identification of causal effects of *manipulable* variables and, finally, they can be tested for validity, albeit indirectly.

Thus, doubts and trepidations concerning the effects of non-manipulable variables and their empirical content should give way to appreciating the important information that these theoretically-defined effects provide. Researchers need not be concerned with the distinction between manipulable and non-manipulative variables, except of course in the design of actual experiments. In the analytical stage, including model specification, identification and estimation, all variables can be treated equally.

MEDIATION ANALYSES AND CROSS-WORLD ASSUMPTIONS

Interviewer: What are your thoughts on decomposing total causal effects into direct and indirect effects?

Pearl: My motivation to engage in effect decomposition was inspired by social scientist Jacques Hagenaars (author of Categorical Longitudinal Data) (Hagenaars, 1993), who convinced me of its importance. Jack considered the distinction between direct and indirect effects to be the key to resolving issues of fairness and discrimination. For example, lawyers explicitly describe sex discrimination as the direct effects of sex on salary, and use counterfactual expressions such as "The candidate income would have been higher had he or she been of different gender" Carson vs Bethlehem Steel Corp., 70 FEP Cases 921, 7th Cir. (1996). Unfortunately, I could not do much with such expressions prior to 1994 for I could not parse the counterfactual relation "Had he or she been." Once I understood how counterfactuals are derived from structural models, and felt comfortable with the algebra that governs counterfactuals, the whole issue of effect decomposition unfolded before my eyes. It came to me as a sudden revelation when I read the legal definition of discrimination: "The central question in any employment-discrimination case is whether the employer would-have taken the same action had the employee been of a different race (age, sex, religion, national origin, etc.) and everything else had been the same" Carson vs Bethlehem Steel Corp., 70 FEP Cases 921, 7th Cir. (1996). All I had to do was to take this sentence and translate it to the algebra of counterfactuals and, Bingo, the definitions of direct and indirect effect came rushing out by themselves, followed by identification conditions, mediation formula, graphical representation, and other goodies (Pearl, 2012ab, 2014b).

I first formalized effect decomposition in 2001 but, in fact, Jamie Robins and Sander Greenland had published a paper on mediation analysis 9 years earlier (Robins & Greenland, 1992) which conceptualized direct and indirect effects in the same counterfactual terms, though not enshrined in mathematical formulas. Remarkably, they ended up concluding that we cannot identify direct and indirect effects, not even from experimental studies. I thought the reason for the difference was that they failed to put the counterfactuals into an equation and were thus prevented from seeing opportunities for identification. Evidently, the reason was more complicated. Jamie felt that it was not scientifically meaningful to make the assumptions needed for identification, for they invoked “cross-world independencies.”

Let me explain. Whenever we say that two unobserved factors are independent, we are likely to make cross-world assumptions. This is because non-confoundedness, the only thing we can test empirically, does not imply ignorability, except for binary treatments. For example, if we say that “the price of beans in China is independent of tomorrow’s traffic in Los Angeles,” we are saying that no matter what the price *p* is, it is independent of all the factors, hidden as well as visible, that may affect tomorrow's traffic level in LA. But we cannot guarantee that if *p* were different than the one actually observed those same factors would still be independent of it. As scientists, we feel comfortable making such assumptions because we cannot find in our theory, or imagination a mechanism that will account for correlation between the two hypothetical factors. Such correlations remain unverifiable however, which make some purists uncomfortable. My perspective remains that such assumptions represent the engine by which we conduct our lives and should not be barred from scientific discourse.

In 2005, my students Chen Avin and Ilya Shpitser came up with a new type of interventions which we called “path specific effects” (Avin, Shpitser, & Pearl, 2005). Instead of interventions that fix variables to constants, their interventions disabled links between variables. Indeed, when we seek to estimate the indirect effect from *X* to *Y* going through a mediator *M*, we need to disable the direct effect. But we cannot disable a direct effect by fixing any variable to a constant. Instead, we sever the direct link and let all other links remain unaltered.

I called the resulting effects “natural,” a term that caught on, because the intervention leaves the units in their natural surroundings (Pearl, 2001). Natural mediation analysis gives us answers to the most important questions that I thought investigators would be concerned with: The percentage of observed effect that could be prevented by disabling the mediating path, and the percentage that would be sustained by the mediating path alone, with the direct effect disabled. These two measures become identical in linear systems but diverge in the presence of interactions, where they capture the difference between the “necessity” and “sufficiency” aspects of mediation (Pearl, 2012b).

The idea of causal mediation analysis has caught on in many fields and has led many researchers to abandon the regression-based analysis that has dominated the social and behavioral sciences since the seminal paper of Baron and Kenny, in 1986 (with 75,000 citations) (Baron & Kenny, 1986). However, the transition has not been easy. Courses in traditional mediation analysis, which treats mediation as a statistical problem are still offered in many universities (Hayes, 2017), and articles in this tradition still appear in conservative publications like the Journal of Structural Equations. Science progresses in turtle steps.

Interviewer: You previously developed Twin Networks. Can you explain when these are useful?

Pearl: Twin Networks were devised to answer one simple question: whether one counterfactual is independent of another, given a third. This type of question was important in the potential outcome framework, where it is called “conditional ignorability” which is required for identification.

Things are totally different in the structural framework. Here, we can establish identification directly from the DAG, using graphical criteria such as “back-door” or “front-door.” These criteria imply “conditional ignorability” without explicitly displaying the counterfactuals involved. So, the twin network is not really needed.

NON-CAUSAL BUT INTERESTING ISSUES

Machine Learning and AI

Interviewer: Machine learning (ML) has scored many achievements in the past decades, and many researchers are aspiring to import ML methods to causal inference problems, as well as harnessing causal inference for achieving human level intelligence. How do you view your work fitting into this context?

Pearl: As I contemplate the success of machine learning I see that the limitations discovered in the causal-inference arena are precisely those that prevent those systems from achieving higher levels of intelligence. The theoretical impediments that prevent us from going from one level of the hierarchy to a higher level also prevent us from reasoning like humans about explanations, regret, fairness, responsibility, and more.

Machine learning is a tool to get us from data to probabilities. But we still need to make two extra steps to go from probabilities into real understanding. 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 (Pearl, 2019a).

In his insightful book Foresight and Understanding (1961), the philosopher Stephen Toulmin argues that the transition away from data-centric thinking is the key to understanding the ancient rivalry between Greek and Babylonian sciences (Toulmin, 1961). 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. Model-blind learning systems may get us to Babylon, but not to Athens.

Mentors

Interviewer: Who do you consider your most important mentors and why?

Pearl: My most important mentors were my high-school teachers who showed us the fun of doing science. My classroom was always decorated with pictures of great scientists on the wall, and I remember putting myself in the shoes of one of those scientists and asking “How would he go about doing this or that?” So, in a sense, I have had many mentors.

I loved Faraday. He never had a formal education. He was a self-educated explorer who used intuition so wisely, that he did not need the aid of math or formal definitions. I remember the day when he invented the first electric motor. When he saw his magnet rotating around an electric wire, he jumped from joy. I danced with him.

I also loved Maxwell because he translated Faraday's intuition into mathematical equations. He showed me that if you translate intuition into mathematics, the mathematics amplifies your intuition and gives you more insight, with which you can decide for example what experiment to try next. Faraday had the intuition of a "field" and Maxwell took it seriously and cast it in 4 differential equations. He then looked at the equations and said: "Wow, they seem to describe a wave. Let's calculate the speed of propagation of that wave. Another Wow! It is the speed of light. Bingo, light must be an electro-magnetic wave." Mind you, this revolutionary discovery came from looking at the form of an equation. I had a terrific kick from this line of thinking, and it still governs much of what I do.

Another mentor was Descartes. I was sick for three days when I learned about his descriptive geometry in high school. The fact that you could do all the constructions of geometry using algebra just blew my mind. I got a high fever and could not get out of bed for three days.

So, I have had many spiritual mentors.

Publishing Technical Reports

Interviewer: Most academics publish in academic journals. You generally publish technical reports, which sometimes come out in journals. Why?

Pearl: This is computer science culture. For a PhD to get a job, you require 8 publications or more in first-rate conferences; journal publications take too long. The technical reports on my website are generally proceedings from such conferences. Most of these were eventually converted to journal articles though the juicy ones were not; these contain heretical ideas or bold criticism of revered leaders in the field (http://bayes.cs.ucla.edu/csl_papers.html). They are waiting for invitation from courageous editors who understand where science is heading.

Bibliography

Angrist, J., & Pischke, J.-S. (2015). *Mastering 'Metrics: The Path from Cause to Effect.* New Jersey: Princeton University Press.

Avin, C., Shpitser, I., & Pearl, J. (2005). Identifiability of path-specific effects. *Proceedings of the Nineteenth International Joint Conference on Artificial Intelligence.* Edinburgh, UK: Morgan Kaufmann Publishers.

Balke, A., & Pearl, J. (1994a). Counterfactual probabilities: Computational methods, bounds, and applications. *Uncertainty in Artificial Intelligence 10* (pp. 46-54). San Mateo: Morgan Kaufmann.

Balke, A., & Pearl, J. (1994b). Probabilistic evaluation of counterfactual queries. *Proceedings of the Twelfth National Conference on Artificial Intelligence* (pp. 230-237). Menlo Park: MIT Press.

Bareinboim, E., & Pearl, J. (2013). A general algorithm for deciding transportability of experimental results. *Journal of Causal Inference, 1*, 107-134.

Bareinboim, E., & Pearl, J. (2016). Causal inference and the data-fusion problem. *Proceedings of the National Academy of Sciences, 113*, 7345-7352.

Baron, R., & Kenny, D. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology, 51*, 1173-1182.

Blalock, H. (1962). Four-variable causal models and partial correlations. *American Journal of Sociology, 68*, 182-194.

Blalock, H. (1964). *Causal Inferences in Nonexperimental Research.* Chapel Hill: University of North Carolina Press.

Bollen, K., & Pearl, J. (2013). Eight myths about causality and structural equation models. In S. Morgan (Ed.), *Handbook of Causal Analysis for Social Research* (pp. 301-328). New York: Springer.

Duncan, O. (1966). Path analysis: sociological examples. *American Journal of Sociology, 72*, 1-16.

Fisher, R. (1926). The arrangement of field experiments. *Journal of the Ministry of Agriculture of Great Britain, 33*, 505-513.

Gibbard, A., & Harper, L. (1981). Counterfactuals and two kinds of expected utility (1976). In W. Harper, R. Stalnaker, & G. Pearce (Eds.), *Ifs* (pp. 153-169). Dordrecht: D. Reidel.

Greenland, S., & Robins, J. (1986). Identifiability, exchangeability, and epidemiological confounding. *International Journal of Epidemiology, 15*, 413-419.

Greenland, S., Pearl, J., & Robins, J. (1999). Causal diagrams for epidemiologic research. *Epidemiology, 10*(1), 37-48.

Haavelmo, T. (1943). The statistical implications of a system of simultaneous equations. *Econometrica, 11*, 1-12.

Hagenaars, J. (1993). *Loglinear Models with Latent Variables.* Newbury Park, CA: Sage Publications.

Halpern, J. (2008). Defaults and normality in causal structures. In G. L. Brewka (Ed.), *Proceedings of the Eleventh International Conference on Principles of Knowledge Representation and Reasoning* (pp. 198-208). San Maeo: Morgan Kaufmann.

Hayes, A. (2017). *Introduction to Mediation, Moderation, and Conditional Process Analysis: A Regression-Based Approach* (2nd Edition ed.). New York: Guilford Press.

Heckman, J., & Pinto, R. (2015). Causal analysis after Haavelmo. *Econometric Theory, 31*(1 (Haavelmo Memorial Issue: Part One)), 115-151.

Holland, P. (1986). Statistics and causal inference. *Journal of the American Statistical Association, 81*, 945-960.

Imbens, G., & Rubin, D. (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences and Introduction.* New York: Cambridge University Press.

Jöreskog, K., & Sörbom, D. (1978). *LISREL IV: Analysis of Linear Structural Relationships by Maximum Likelihood.* Chicago: International Educational Services.

Morgan, S., & Winship, C. (2007). *Counterfactuals and Causal Inference: Methods and Principles for Social Research (Analytical Methods for Social Research).* New York: Cambridge University Press.

Morgan, S., & Winship, C. (2015). *Counterfactuals and Causal Inference: Methods and Principles for Social Research (Analytical Methods for Social Research)* (2nd ed.). New York: Cambridge University Press.

Muthen, B. (1987). Response to Freedman's critique of path analysis: Improve credibility by better methodological training. *Journal of Educational Statistics, 12*, 178-184.

Neyman, J. (1923). Sur les applications de la thar des probabilities aux experiences Agaricales: Essay des principle. *Statistical Science, 5*, 463-472.

Pearl, J. (1986). Fusion, propagation, and structuring in belief networks. *Artificial Intelligence, 29*, 241-288.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems.* San Mateo: Morgan Kaufmann.

Pearl, J. (1993). Graphical Models, Causality, and Intervention. *Statistical Science, 8*, 266-269.

Pearl, J. (1994). A probabilistic calculus of actions. In R. de Mantaras, & D. Poole (Ed.), *Uncertainty in Artificial Intelligence 10* (pp. 454-462). San Mateo: Morgan Kaufmann.

Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika, 82*, 669-710.

Pearl, J. (2001). Direct and indirect effects. *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence* (pp. 411-420). San Francisco: Morgan Kaufmann.

Pearl, J. (2009). *Myth, Confusion, and Science in Causal Analysis.* Retrieved from UCLA Computer Science Department, Technical Report R-348, May 2009: http://ftp.cs.ucla.edu/pub/stat_ser/r348-warning.pdf

Pearl, J. (2012a). The Causal Mediation Formula -- A Guide to the Assessment of Pathways and Mechanisms. *Prevention Science, 13*, 426-436.

Pearl, J. (2012b). The Mediation Formula: A guide to the assessment of causal pathways in nonlinear models. In C. Berzuini, P. Dawid, & L. Bernardinelli (Eds.), *Causality: Statistical Perspectives and Applications* (pp. 151-179). Chichester, UK: John Wiley and Sons, Ltd.

Pearl, J. (2014a). Graphoids over counterfactuals. *Journal of Causal Inference, 2*, 243-248.

Pearl, J. (2014b). Reply to Commentary by Imai, Keele, Tingley, and Yamamoto Concerning Causal Mediation Analysis. *Psychological Methods, 19*(4), 488-492.

Pearl, J. (2015a). Trygve Haavelmo and the emergence of causal calculus. *Econometric Theory, 31*(1), 152-179.

Pearl, J. (2015b). Causes of effects and effects of causes. *Journal of Sociological Methods and Research, 44*, 149-164.

Pearl, J. (2019a). The seven tools of causal inference, with reflections on machine learning. *Communications of Association for Computing Machinery, 62*, 54-60.

Pearl, J. (2019b). On the intepretation of do(x). *Journal of Causal Inference, 7*, online.

Pearl, J., & Bareinboim, E. (2014). External validity: From do-calculus to transportability across populations. *Statistical Science, 29*, 579-595.

Pearl, J., & Mackenzie, D. (2018). *The Book of Why: The New Science of Cause and Effect.* New York: Basic Books.

Pearl, J., & Robins, J. (1995). Probabilistic evaluation of sequential plans from causal models with hidden variables. In P. Besnard, & S. Hanks (Ed.), *Uncertainty in Artificial Intelligence 11* (pp. 444-453). San Francisco: Morgan Kaufmann.

Robins, J. (1986). A new approach to causal inference in mortality studies with a sustained exposure period -- applications to control of the healthy workers survivor effect. *Mathematical Modeling, 7*, 1393-1512.

Robins, J., & Greenland, S. (1992). Identifiability and exchangeability for direct and indirect effects. *Epidemiology, 3*, 143-155.

Rubin, D. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology, 66*, 688-701.

Rubin, D. (1980). Randomization analysis of experimental data: The Fisher randomization test comment. *Journal of the American Statistical Association, 75*, 591-593.

Rubin, D. (2009). Author's reply: Should observational studies be designed to allow lack of balance in covariate distributions across treatment group? *Statistics in Medicine, 28*, 1420-1423.

Shpitser, I., & Pearl, J. (2006). Identification of joint interventional distributions in recursive semi-Markovian causal models. *Proceedings of the Twenty- First National Conference on Artificial Intelligence* (pp. 1219-1226). Menlo Park: AAAI Press.

Sobel, M. (2009). Causal inference in randomized and non-randomized studies: the definition, identification, and estimation of causal parameters. In R. Millsap, & Mayadeu-Olivares (Eds.), *Quantitative Methods in Psychology* (pp. 3-22). Los Angeles: SAGE.

Spirtes, P., Glymour, C., & Scheines, R. (1993). *Causation, Prediction, and Search.* New York: Springer-Verlag.

Stalnaker, R. (1980). Letter to David Lewis (May 21, 1972). In W. Harper, R. Stalnaker, & G. Pearce (Eds.), *Ifs* (Vol. 15, pp. 151-152). Dordrecht: D. Reidel.

Strotz, R., & Wold, H. (1960). Recursive versus nonrecursive systems: An attempt at synthesis. *Econometrica, 28*, 417-427.

Toulmin, S. (1961). *Forecast and Understanding.* Indiana: University Press.

Wright, P. (1928). *The Tariff on Animal and Vegetable Oils.* New York: The MacMillan Company.

Wright, S. (1921). Correlation and causation. *Journal of Agricultural Research, 20*, 557-585.