TECHNICAL REPORT
R-475
September 2017

Theoretical Impediments to Machine Learning
With Seven Sparks from the Causal Revolution

Judea Pearl, University of California, Los Angeles
September 2017

Abstract

Current machine learning systems operate, almost exclusively, in a statistical, or model-blind mode, which entails severe theoretical limits on their power and performance. Such systems cannot reason about interventions and retrospection and, therefore, cannot serve as the basis for strong AI. To achieve human level intelligence, learning machines need the guidance of a model of reality, similar to the ones used in causal inference. To demonstrate the essential role of such models, I will present a summary of seven tasks which are beyond reach of current machine learning systems and which have been accomplished using the tools of causal inference.

Scientific Background

If we examine the information that drives machine learning today, we find that it is almost entirely statistical. In other words, learning machines improve their performance by optimizing parameters over a stream of sensory inputs received from the environment. It is a slow process, analogous in many respects to the survival-of-the-fittest 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 one thousand years. What humans possessed that other species lacked was a mental representation, a blue-print of their environment which they could manipulate at will to *imagine* alternative hypothetical environments for planning and learning. Anthropologists like N. Harari, and S. Mithen are in general agreement that the decisive ingredient that gave our Homo sapiens ancestors the ability to achieve global dominion, about 40,000 years ago, was their ability to sketch and store a mental representation of their environment, interrogate that representation, distort it by mental acts of imagination and finally answer "What if?" kind of questions. Examples are interventional questions: "What if I act?" and retrospective or explanatory questions: "What if I had acted differently?" No learning machine in operation today can answer such questions about interventions not taken before, say, "What if we ban cigarettes." Moreover, most learning machines today do not provide a representation from which the answers to such questions can be derived.

I postulate that the major impediment to achieving accelerated learning speeds as well as human level performance should be overcome by removing these barriers and equipping learning machines with causal reasoning tools. This postulate would have been speculative twenty years ago, prior to the mathematization of counterfactuals. Not so today.

Advances in graphical and structural models have made counterfactuals computationally manageable and thus rendered model-driven reasoning a more promising direction on which to base strong AI. In the next section, I will describe the impediments facing machine learning systems using a three-level hierarchy that governs inferences in causal reasoning. The final section summarizes how these impediments were circumvented using modern tools of causal inference.

The Three Layer Causal Hierarchy

Level (Symbol)	Typical Activity	Typical Questions	Examples
1. Association $P(y\|x)$	Seeing	What is? How would seeing X change my belief in$Y$?	What does a symptom tell me about a disease? What does a survey tell us about the election results?
2. Intervention $P(y\|do(x), z)$	Doing	What if? What if I do X?	What if I take aspirin, will my headache be cured? What if we ban cigarettes?
3. Counterfactuals $P(y_x\|x', y')$	Imagining, Retrospection	Why? Was it X that caused Y? What if I had acted differently?	Was it the aspirin that stopped my headache? Would Kennedy be alive had Oswald not shot him? What if I had not been smoking the past 2 years?

Figure 1: The ladder of causation

An extremely useful insight unveiled by the logic of causal reasoning is the existence of a sharp classification of causal information, in terms of the kind of questions that each class is capable of answering. The classification forms a 3-level hierarchy in the sense that questions at level i $(i = 1, 2, 3)$ can only be answered if information from level j $(j \geq i)$ is available.

Figure 1 shows the 3-level hierarchy, together with the characteristic questions that can be answered at each level. The levels are titled 1. Association, 2. Intervention, and 3. Counterfactual. The names of these layers were chosen to emphasize their usage. We call the first level Association, because it invokes purely statistical relationships, defined by the naked data.[1] For instance, observing a customer who buys toothpaste makes it more likely that he/she buys floss; such association can be inferred directly from the observed data using conditional expectation. Questions at this layer, because they require no causal information, are placed at the bottom level on the hierarchy. The second level, Intervention, ranks higher than Association because it involves not just seeing what is, but changing what we see. A typical question at this level would be: What happens if we double the price? Such questions cannot be answered from sales data alone, because they involve a change in customers behavior, in reaction to the new pricing. These choices

[1] Other names used for inferences at this layer are: “model-free,” “model-blind,” “black-box,” or “data-centric.” Darwiche (2017) used “function-fitting,” for it amounts to fitting data by a complex function defined by the neural network architecture.

may differ substantially from those taken in previous price-raising situations. (Unless we replicate precisely the market conditions that existed when the price reached double its current value.) Finally, the top level is called Counterfactuals, a term that goes back to the philosophers David Hume and John Stewart Mill, and which has been given computer-friendly semantics in the past two decades. A typical question in the counterfactual category is "What if I had acted differently," thus necessitating retrospective reasoning.

Counterfactuals are placed at the top of the hierarchy because they subsume interventional and associational questions. If we have a model that can answer counterfactual queries, we can also answer questions about interventions and observations. For example, the interventional question, What will happen if we double the price? can be answered by asking the counterfactual question: What would happen had the price been twice its current value? Likewise, associational questions can be answered once we can answer interventional questions; we simply ignore the action part and let observations take over. The translation does not work in the opposite direction. Interventional questions cannot be answered from purely observational information (i.e., from statistical data alone). No counterfactual question involving retrospection can be answered from purely interventional information, such as that acquired from controlled experiments; we cannot re-run an experiment on subjects who were treated with a drug and see how they behave had they not given the drug. The hierarchy is therefore directional, with the top level being the most powerful one.

Counterfactuals are the building blocks of scientific thinking as well as legal and moral reasoning. In civil court, for example, the defendant is considered to be the culprit of an injury if, *but for* the defendant's action, it is more likely than not that the injury would not have occurred. The computational meaning of *but for* calls for comparing the real world to an alternative world in which the defendant action did not take place.

Each layer in the hierarchy has a syntactic signature that characterizes the the sentences admitted into that layer. For example, the association layer is characterized by conditional probability sentences, e.g., $P(y|x) = p$ stating that: the probability of event $Y = y$ given that we observed event $X = x$ is equal to p. In large systems, such evidential sentences can be computed efficiently using Bayesian Networks, or any of the neural networks that support deep-learning systems.

At the interventional layer we find sentences of the type $P(y|do(x), z)$, which denotes "The probability of event $Y = y$ given that we intervene and set the value of X to x and subsequently observe event $Z = z$. Such expressions can be estimated experimentally from randomized trials or analytically using Causal Bayesian Networks (Pearl, 2000, Chapter 3).

Finally, at the counterfactual level, we have expressions of the type $P(y_x|x', y')$ which stand for "The probability of event $Y = y$ had X been x, given that we actually observed X to be x' and and Y to be y'. For example, the probability that Joe's salary would be y had he finished college, given that his actual salary is y' and that he had only two years of college." Such sentences can be computed only when we possess functional or Structural Equation models, or properties of such models (Pearl, 2000, Chapter 7).

This hierarchy, and the formal restrictions it entails, explains why statistics-based machine learning systems are prevented from reasoning about actions, experiments and explanations. It also informs us what extra-statistical information is needed, and in what format, in order to support those modes of reasoning.

The Seven Pillars of the Causal Revolution
(or What you can do with a causal model that you could not do without?)

Consider the following five questions:

- How effective is a given treatment in preventing a disease?
- Was it the new tax break that caused our sales to go up?
- What is the annual health-care costs attributed to obesity?
- Can hiring records prove an employer guilty of sex discrimination?
- I am about to quit my gob, but should I?

The common feature of these questions is that they are concerned with cause-and-effect relationships. We can recognize them through words such as ”preventing”, ”cause,” “attributed to,” “discrimination,” and “should I.” Such words are common in everyday language, and our society constantly demands answers to such questions. Yet, until very recently science gave us no means even to articulate them, let alone answer them. Unlike the rules of geometry, mechanics, optics or probabilities, the rules of cause and effect have been denied the benefits of mathematical analysis.

To appreciate the extent of this denial, readers would be stunned to know that only a few decades ago scientists were unable to write down a mathematical equation for the obvious fact that “mud does not cause rain.” Even today, only the top echelon of the scientific community can write such an equation and formally distinguish “mud causes rain” from “ rain causes mud.” And you would probably be even more surprised to discover that your favorite college professor is not among them.

Things have changed dramatically in the past three decades, A powerful yet transparent mathematical language has been developed for managing causes and effects, accompanied by a set of tools that turn causal analysis into a mathematical game, not unlike solving algebraic equations, or finding proofs in high-school geometry. These tools permit us to express causal questions formally codify our existing knowledge in both diagrammatic and algebraeid forms, and then leverage our data to estimate the answers. Moreover, the theory warns us when the state of existing knowledge or the available data are insufficient to answer our questions; and then suggests additional sources of knowledge or data to make the questions answerable.

I call this transformation “The Causal Revolution,” (Pearl and Mackenzie, 2018, forthcoming) and the mathematical framework that led to it I call “Structural Causal Models (SCM).”

The SCM deploys three parts

1. Graphical models,
2. Structural equations, and
3. Counterfactual and interventional logic

Graphical models serve as a language for representing what we know about the world, counterfactuals help us to articulate what we want to know, while structural equations serve to tie the two together in a solid semantics.

Next we provide a bird's eye view of the seven most significant accomplishments of the SCM framework and discuss the unique contribution that each pillar brings to the art of automated reasoning.

Pillar 1: Encoding Causal Assumptions – Transparency and Testability

Advances in graphical models made it possible to encode causal assumptions in a compact format, maintaining transparency and testability. Transparency enables analysts to discern whether the assumptions encoded are plausible (on scientific grounds), and whether additional assumptions are warranted. Testability permits us (be it an analyst or a machine) to determine whether the assumptions encoded are compatible with the available data and, if not, identify those that need repair. This is facilitated through a graphical criterion called d-separation, which constitutes the fundamental connection between causes and probabilities. It tells us, for any given pattern of paths in the model, what pattern of dependencies we should expect in the data (Pearl, 1988).

Pillar 2: Do-calculus and the control of confounding

Confounding, the major obstacle to drawing causal inference from data, had been demystified and totally "deconfounded" using a graphical criterion called "back-door." In particular, the task of selecting an appropriate set of covariates for control of confounding has been reduced to a simple "roadblocks" puzzle manageable by a simple algorithm (Pearl, 1993).

In cases where the "back-door" criterion does not hold, a symbolic engine has been developed, called *do-calculus*, which predicts the effect of policy interventions whenever feasible, and exits with failure whenever predictions cannot be ascertained with the specified assumptions (Pearl, 1995; Tian and Pearl, 2002; Shpitser and Pearl, 2008).

Pillar 3: The Algorithmization of Counterfactuals

Counterfactual analysis deals with behavior of specific individuals, identified by a distinct set of characteristics, For example, given that Joe's salary is $Y = y$, and that he attended $X = x$ years of college, what would Joe's salary be had he had one more year of education.

One of the crown achievements of the Causal Revolution has been to formulate counterfactual reasoning within the graphical representation, the very representation researchers use to encode scientific knowledge. Every structural equation model determines the truth value of every counterfactual sentence. Therefore, we can determine analytically if the probability that such a sentence is true is estimable from experimental or observational studies, or combination thereof (Balke and Pearl, 1994; Pearl, 2000, Chapter 7).

Of special interest in causal discourse are counterfactual questions concerning "causes of effects," as opposed to "effects of causes." For example, how likely it is that Joe's swimming exercise was a necessary (or sufficient) cause of Joe's death (Pearl, 2015).

Pillar 4: Mediation Analysis and the Assessment of Direct and Indirect Effects

Mediation analysis concerns the mechanisms that transmit changes from a cause to its effects. The detection of intermediate mechanism is essential for generating explanations and counterfactual logic must be invoked to facilitate this detection. The graphical representation of counterfactuals enables us to define direct and indirect effects and to decide when these effects are estimable from data, or experiments (Robins and Greenland, 1992; Pearl, 2001; VanderWeele, 2015).

Pillar 5: External Validity and Sample Selection Bias

The validity of every experimental study is challenged by disparities between the experimental and implementational setups. A machine trained in one environment cannot be expected to perform well when environmental conditions change, unless the changes are localized and identified. The *do*-calculus discussed above now offers a complete methodology for overcoming this source of bias. It can be used both for re-adjusting learned policies to circumvent environmental changes and for controlling bias due to non-representative samples (Bareinboim and Pearl, 2016).

Pillar 6: Missing Data

Problems of missing data plague every branch of experimental science. Respondents do not answer every item on a questionnaire, sensors fade as environmental conditions change, and patients often drop from a clinical study for unknown reasons. The rich literature on this problem is wedded to the model-blind paradigm of statistical analysis. Using causal models of the missingness process we can now formalize the conditions under which causal and probabilistic relationships can be recovered from incomplete data and, whenever the conditions are satisfied, produce a consistent estimate of the desired relationship (Mohan and Pearl, 2017).

Pillar 7: Causal Discovery

The d-separation criterion described above enables us to detect and enumerate the testable implications of a given causal model. This opens the possibility of inferring, with very mild assumptions, the set of models that are compatible with the data, and to represent this set compactly. Systematic searches have been developed which, in certain circumstances, can prune the set of compatible models significantly to the point where causal queries can be estimated directly from that set (Spirtes et al., 2000; Peters et al., 2017, in press).

Conclusions

Model-blind approaches to AI have intrinsic limitations on the cognitive tasks that they can perform. We have described some of these tasks and demonstrated how they can be accomplished in the framework of causal reasoning. Data science is only as much of a science as it facilitates the interpretation of data – a two-body task involving both data and reality. Data alone are hardly a science, regardless how plenty they get and how skillfully they are summarized.

References

BALKE, A. and PEARL, J. (1994). Probabilistic evaluation of counterfactual queries. In *Proceedings of the Twelfth National Conference on Artificial Intelligence*, vol. I. MIT Press, Menlo Park, CA, 230–237.

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

DARWICHE, A. (2017). Human-level intelligence or animal-like abilities? Tech. rep., Department of Computer Science, University of California, Los Angeles, CA. ArXiv:1707.04327.

MOHAN, K. and PEARL, J. (2017). Graphical models for processing missing data. Tech. Rep. R-473, <http://ftp.cs.ucla.edu/pub/stat_ser/r473.pdf>, Department of Computer Science, University of California, Los Angeles, CA. Submitted.

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

PEARL, J. (1993). Comment: Graphical models, causality, and intervention. *Statistical Science* **8** 266–269.

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

PEARL, J. (2000). *Causality: Models, Reasoning, and Inference*. Cambridge University Press, New York. 2nd edition, 2009.

PEARL, J. (2001). Direct and indirect effects. In *Uncertainty in Artificial Intelligence, Proceedings of the Seventeenth Conference*. Morgan Kaufmann, San Francisco, CA, 411–420.

PEARL, J. (2015). Causes of effects and effects of causes. *Journal of Sociological Methods and Research* **44** 149–164.

PEARL, J. and MACKENZIE, D. (2018, forthcoming). *The Book of Why: The New Science of Cause and Effect*. Basic Books, New York.

PETERS, J., JANZING, D. and SCHÖLKOPF, B. (2017, in press). *Elements of Causal Inference – Foundations and Learning Algorithms*. The MIT Press, Cambridge, MA.

ROBINS, J. and GREENLAND, S. (1992). Identifiability and exchangeability for direct and indirect effects. *Epidemiology* **3** 143–155.

SHPITSER, I. and PEARL, J. (2008). Complete identification methods for the causal hierarchy. *Journal of Machine Learning Research* **9** 1941–1979.

SPIRTES, P., GLYMOUR, C. and SCHEINES, R. (2000). *Causation, Prediction, and Search*. 2nd ed. MIT Press, Cambridge, MA.

TIAN, J. and PEARL, J. (2002). A general identification condition for causal effects. In *Proceedings of the Eighteenth National Conference on Artificial Intelligence*. AAAI Press/The MIT Press, Menlo Park, CA, 567–573.

VANDERWEELE, T. (2015). *Explanation in Causal Inference: Methods for Mediation and Interaction*. Oxford University Press, New York.