Communications of the ACM

Research and Advances

Artificial Intelligence and Machine Learning

The Seven Tools of Causal Inference, with Reflections on Machine Learning

The kind of causal inference seen in natural human thought can be "algorithmitized" to help produce human-level machine intelligence.

By Judea Pearl

Posted Mar 1 2019

The dramatic success in machine learning has led to an explosion of artificial intelligence (AI) applications and increasing expectations for autonomous systems that exhibit human-level intelligence. These expectations have, however, met with fundamental obstacles that cut across many application areas. One such obstacle is adaptability, or robustness. Machine learning researchers have noted current systems lack the ability to recognize or react to new circumstances they have not been specifically programmed or trained for. Intensive theoretical and experimental efforts toward "transfer learning," "domain adaptation," and "lifelong learning"[4] are reflective of this obstacle.

Back to Top

Key Insights

- Data science is a two-body problem, connecting data and reality, including the forces behind the data.
- Data science is the art of interpreting reality in the light of data, not a mirror through which data sees itself from different angles.
- The ladder of causation is the double helix of causal thinking, defining what can and cannot be learned about actions and about worlds that could have been.

Another obstacle is "explainability," or that "machine learning models remain mostly black boxes"[26] unable to explain the reasons behind their predictions or recommendations, thus eroding users' trust and impeding diagnosis and repair; see Hutson[8] and Marcus.[11] A third obstacle concerns the lack of understanding of cause-effect connections. This hallmark of human cognition[10,23] is, in my view, a necessary (though not sufficient) ingredient for achieving human-level intelligence. This ingredient should allow computer systems to choreograph a parsimonious and modular representation of their environment, interrogate that representation, distort it through acts of

imagination, and finally answer "What if?" kinds of questions. Examples include interventional questions: "What if I make it happen?" and retrospective or explanatory questions: "What if I had acted differently?" or "What if my flight had not been late?" Such questions cannot be articulated, let alone answered by systems that operate in purely statistical mode, as do most learning machines today. In this article, I show that all three obstacles can be overcome using causal modeling tools, in particular, causal diagrams and their associated logic. Central to the development of these tools are advances in graphical and structural models that have made counterfactuals computationally manageable and thus rendered causal reasoning a viable component in support of strong AI.

In the next section, I describe a three-level hierarchy that restricts and governs inferences in causal reasoning. The final section summarizes how traditional impediments are circumvented through modern tools of causal inference. In particular, I present seven tasks that are beyond the reach of "associational" learning systems and have been (and can be) accomplished only through the tools of causal modeling.

Back to Top

The Three-Level Causal Hierarchy

A useful insight brought to light through the theory of causal models is the classification of causal information in terms of the kind of questions each class is capable of answering. The classification forms a three-level hierarchy in the sense that questions at level i (i = 1, 2, 3) can be answered only if information from level j ($j > i$) is available.

Figure 1 outlines the three-level hierarchy, together with the characteristic questions that can be answered at each level. I call the levels 1. Association, 2. Intervention, and 3. Counter-factual, to match their usage. I call the first level Association because it invokes purely statistical relationships, defined by the naked data.[a] For instance, observing a customer who buys toothpaste makes it more likely that this customer will also buy floss; such associations can be inferred directly from the observed data using standard conditional probabilities and conditional expectation.[15] Questions at this layer, because they require no causal information, are placed at the bottom level in the hierarchy. Answering them is the hallmark of current machine learning methods.[4] 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 will happen if we double the price? Such a question cannot be answered from sales data alone, as it involves a change in customers' choices in reaction to the new pricing. These choices 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 invokes Counterfactuals, a mode of reasoning that goes back to the philosophers David Hume and John Stuart Mill and that has been given computer-friendly semantics in the past two decades.[1,18] A typical question in the counterfactual category is: "What if I had acted differently?" thus necessitating retrospective reasoning.

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, Intervening	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 two years?

Figure 1. The causal hierarchy. Questions at level 1 can be answered only if information from level i or higher is available.

I place Counterfactuals 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 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, from statistical data alone. No counterfactual question involving retrospection can be answered from purely interventional information, as with that acquired from controlled experiments; we cannot re-run an experiment on human subjects who were treated with a drug and see how they might behave had they not been 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 of legal and moral reasoning. For example, in civil court, a defendant is considered responsible for an injury if, but for the defendant's action, it is more likely than not 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's action did not take place.

Each layer in the hierarchy has a syntactic signature that characterizes the sentences admitted into that layer. For example, the Association layer is characterized by conditional probability sentences, as in $P(y|x) = p$, stating that: The probability of event $Y=y$, given that we observed event $X = x$ is equal to p. In large systems, such evidentiary sentences can be computed efficiently through Bayesian networks or any number of machine learning techniques.

At the Intervention layer, we deal with sentences of the type $P(y|do(x), z)$ that denote "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.[18] A child learns the effects of interventions through playful manipulation of the environment (usually in a deterministic playground), and AI planners obtain interventional knowledge by exercising admissible sets of actions. Interventional expressions cannot be inferred from passive observations alone, regardless of how big the data.

Finally, at the Counterfactual level, we deal with expressions of the type $P(yx\,|x',y')$ that stand for "The probability that event $Y = y$ would be observed had X been x, given that we actually observed X

to be x 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 the model is based on functional relations or structural.[18]

This three-level hierarchy, and the formal restrictions it entails, explains why machine learning systems, based only on associations, are prevented from reasoning about (novel) actions, experiments, and causal explanations.[b]

Back to Top

Questions Answered with a Causal Model

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 annual health-care costs are attributed to obesity?;
- Can hiring records prove an employer guilty of sex discrimination?; and
- I am about to quit my job, but should I?

The common feature of these questions concerns cause-and-effect relationships. We recognize them through such words as "preventing," "cause," "attributed to," "discrimination," and "should I." Such words are common in everyday language, and modern 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 likely be stunned to learn 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."

These impediments have changed dramatically in the past three decades; for example, a mathematical language has been developed for managing causes and effects, accompanied by a set of tools that turn causal analysis into a mathematical game, like solving algebraic equations or finding proofs in high-school geometry. These tools permit scientists to express causal questions formally, codify their existing knowledge in both diagrammatic and algebraic forms, and then leverage data to estimate the answers. Moreover, the theory warns them when the state of existing knowledge or the available data is insufficient to answer their questions and then suggests additional sources of knowledge or data to make the questions answerable.

The development of the tools has had a transformative impact on all data-intensive sciences, especially social science and epidemiology, in which causal diagrams have become a second language.[14,34] In these disciplines, causal diagrams have helped scientists extract causal relations from associations and deconstruct paradoxes that have baffled researchers for decades.[23,25]

I call the mathematical framework that led to this transformation "structural causal models" (SCM), which consists of three parts: graphical models, structural equations, and counterfactual and interventional logic. Graphical models serve as a language for representing what agents know about

the world. Counterfactuals help them articulate what they wish to know. And structural equations serve to tie the two together in a solid semantics.

Figure 2 illustrates the operation of SCM in the form of an inference engine. The engine accepts three inputs—Assumptions, Queries, and Data—and produces three outputs—Estimand, Estimate, and Fit indices. The Estimand (E_S) is a mathematical formula that, based on the Assumptions, provides a recipe for answering the Query from any hypothetical data, whenever it is available. After receiving the data, the engine uses the Estimand to produce an actual Estimate ($\hat{E}_S$) for the answer, along with statistical estimates of the confidence in that answer, reflecting the limited size of the dataset, as well as possible measurement errors or missing data. Finally, the engine produces a list of "fit indices" that measure how compatible the data is with the Assumptions conveyed by the model.

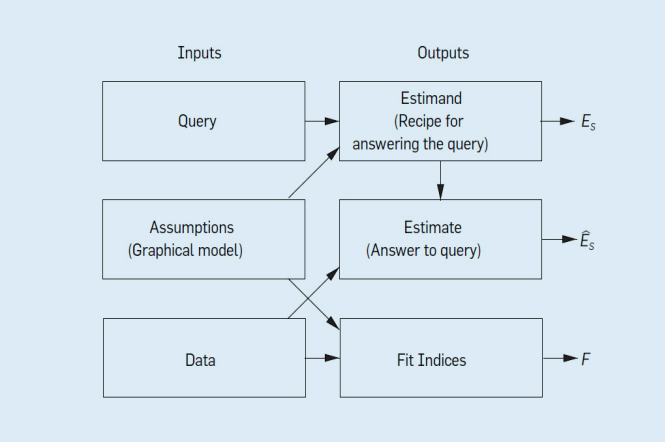

Figure 2. How the SCM "inference engine" combines data with a causal model (or assumptions) to produce answers to queries of interest.

To exemplify these operations, assume our Query stands for the causal effect of X (taking a drug) on Y (recovery), written as $Q = P(Y|do(X))$. Let the modeling assumptions be encoded (see Figure 3), where Z is a third variable (say, Gender) affecting both X and Y. Finally, let the data be sampled at random from a joint distribution $P(X, Y, Z)$. The Estimand (E_S) derived by the engine (automatically using Tool 2, as discussed in the next section) will be the formula $E_S = \Sigma z\, P(Y|X, Z)P(Z)$, which defines a procedure of estimation. It calls for estimating the gender-specific conditional distributions $P(Y|X, Z)$ for males and females, weighing them by the probability $P(Z)$ of membership in each gender, then taking the average. Note the Estimand E_S defines a property of $P(X, Y, Z)$ that, if properly estimated, would provide a correct answer to our Query. The answer itself, the Estimate $\hat{E}_S$, can be produced through any number of techniques that produce a consistent estimate of ES from finite samples of $P(X, Y, Z)$. For example, the sample average (of Y) over all cases satisfying the specified X and Z conditions would be a consistent estimate. But more-efficient estimation techniques can be devised to overcome data sparsity.[28] This task of estimating statistical relationships from sparse data is where deep learning techniques excel, and where they are often employed.[33]

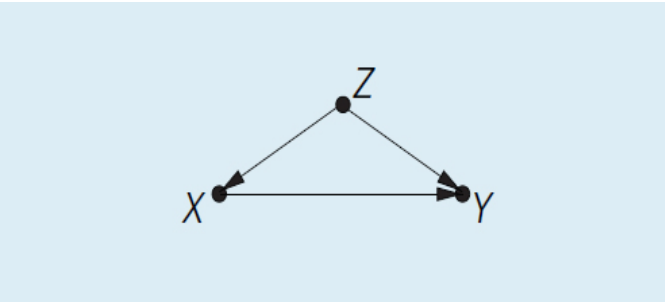

Figure 3. Graphical model depicting causal assumptions about three variables; the task is to estimate the causal effect of *X* on *Y* from non-experimental data on {*X*, *Y*, *Z*}.

Finally, the Fit Index for our example in Figure 3 will be NULL; that is, after examining the structure of the graph in Figure 3, the engine should conclude (using Tool 1, as discussed in the next section) that the assumptions encoded lack testable implications. Therefore, the veracity of the resultant estimate must lean entirely on the assumptions encoded in the arrows of Figure 3, so neither refutation nor corroboration can be obtained from the data.[c]

The same procedure applies to more sophisticated queries, as in, say, the counterfactual query $Q = P(y_x \mid x', y')$ discussed earlier. We may also permit some of the data to arrive from controlled experiments that would take the form $P(V|do(W))$ in case W is the controlled variable. The role of the Estimand would remain that of converting the Query into the syntactic form involving the available data and then guiding the choice of the estimation technique to ensure unbiased estimates. The conversion task is not always feasible, in which case the Query is declared "non-identifiable," and the engine should exit with FAILURE. Fortunately, efficient and complete algorithms have been developed to decide identifiability and produce Estimands for a variety of counterfactual queries and a variety of data types.[3,30,32]

I next provide a bird's-eye view of seven tasks accomplished through the SCM framework and the tools used in each task and discuss the unique contribution each tool brings to the art of automated reasoning.

Tool 1. Encoding causal assumptions: Transparency and testability. The task of encoding assumptions in a compact and usable form is not a trivial matter once an analyst takes seriously the requirement of transparency and testability.[d] Transparency enables analysts to discern whether the assumptions encoded are plausible (on scientific grounds) or whether additional assumptions are warranted. Testability permits us (whether analyst or machine) to determine whether the assumptions encoded are compatible with the available data and, if not, identify those that need repair.

Advances in graphical models have made compact encoding feasible. Their transparency stems naturally from the fact that all assumptions are encoded qualitatively in graphical form, mirroring the way researchers perceive cause-effect relationships in the domain; judgments of counterfactual or statistical dependencies are not required, since such dependencies can be read off the structure of the graph.[18] Testability is facilitated through a graphical criterion called *d*-separation that provides 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 to find in the data.[15]

Tool 2. *Do*-calculus and the control of confounding. Confounding, or the presence of unobserved causes of two or more variables, long considered *the* major obstacle to drawing causal inference from data, has been demystified and "deconfounded" through a graphical criterion called "backdoor." In particular, the task of selecting an appropriate set of covariates to control for confounding has been reduced to a simple "roadblocks" puzzle manageable through a simple algorithm.[16]

For models where the backdoor criterion does not hold, a symbolic engine is available, called "*do*-calculus," that predicts the effect of policy interventions whenever feasible and exits with failure whenever predictions cannot be ascertained on the basis of the specified assumptions.[3,17,30,32]

Tool 3. The algorithmitization 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 went $X = x$ years to college, what would Joe's salary be had he had one more year of education?

One of the crowning achievements of contemporary work on causality has been to formalize 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, an algorithm can determine if the probability of the sentence is estimable from experimental or observational studies, or a combination thereof.[1,18,30]

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.[7,20]

Tool 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 identification of such an intermediate mechanism is essential for generating explanations, and counterfactual analysis must be invoked to facilitate this identification. The logic of counterfactuals and their graphical representation have spawned algorithms for estimating direct and indirect effects from data or experiments.[19,27,34] A typical query computable through these algorithms is: What fraction of the effect of X on Y is mediated by variable Z?

Tool 5. Adaptability, external validity, and sample selection bias. The validity of every experimental study is challenged by disparities between the experimental and the intended 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. This problem, and its various manifestations, are well-recognized by AI researchers, and enterprises (such as "domain adaptation," "transfer learning," "life-long learning," and "explainable AI")[4] are just some of the subtasks identified by researchers and funding agencies in an attempt to alleviate the general problem of robustness. Unfortunately, the problem of robustness, in its broadest form, requires a causal model of the environment and cannot be properly addressed at the level of Association. Associations alone cannot identify the mechanisms responsible for the changes that occurred,[22] the reason being that surface changes in observed associations do not uniquely identify the underlying mechanism responsible for the change. The *do*-calculus discussed earlier now offers a complete methodology for overcoming bias due to environmental changes. It can be used for both for readjusting learned policies to circumvent environmental changes and for controlling disparities between nonrepresentative samples and a target population.[3] It can also be used in the context of reinforcement learning to evaluate policies that invoke new actions, beyond those used in training.[35]

Unlike the rules of geometry, mechanics, optics, or probabilities, the rules of cause and effect have been denied the benefits of mathematical analysis.

Tool 6. Recovering from missing data. Problems due to missing data plague every branch of experimental science. Respondents do not answer every item on a questionnaire, sensors malfunction as weather conditions worsen, and patients often drop from a clinical study for

unknown reasons. The rich literature on this problem is wedded to a model-free paradigm of associational analysis and, accordingly, is severely limited to situations where "missingness" occurs at random; that is, independent of values taken by other variables in the model.[6] 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.[12,13]

Tool 7. Causal discovery. The *d*-separation criterion described earlier enables machines to detect and enumerate the testable implications of a given causal model. This opens the possibility of inferring, with mild assumptions, the set of models that are compatible with the data and to represent this set compactly. Systematic searches have been developed that, in certain circumstances, can prune the set of compatible models significantly to the point where causal queries can be estimated directly from that set.[9,18,24,31]

Alternatively, Shimizu et al.[29] proposed a method for discovering causal directionality based on functional decomposition.[24] The idea is that in a linear model $X \rightarrow Y$ with non-Gaussian noise, $P(y)$ is a convolution of two non-Gaussian distributions and would be, figuratively speaking, "more Gaussian" than $P(x)$. The relation of "more Gaussian than" can be given precise numerical measure and used to infer directionality of certain arrows.

Tian and Pearl[32] developed yet another method of causal discovery based on the detection of "shocks," or spontaneous local changes in the environment that act like "nature's interventions," and unveil causal directionality toward the consequences of those shocks.

Back to Top

Conclusion

I have argued that causal reasoning is an indispensable component of human thought that should be formalized and algorithmitized toward achieving human-level machine intelligence. I have explicated some of the impediments toward that goal in the form of a three-level hierarchy and showed that inference to level 2 and level 3 requires a causal model of one's environment. I have described seven cognitive tasks that require tools from these two levels of inference and demonstrated how they can be accomplished in the SCM framework.

It is important for researchers to note that the models used in accomplishing these tasks are structural (or conceptual) and require no commitment to a particular form of the distributions involved. On the other hand, the validity of all inferences depends critically on the veracity of the assumed structure. If the true structure differs from the one assumed, and the data fits both equally well, substantial errors may result that can sometimes be assessed through a sensitivity analysis.

It is also important for them to keep in mind that the theoretical limitations of model-free machine learning do not apply to tasks of prediction, diagnosis, and recognition, where interventions and counterfactuals assume a secondary role.

However, the model-assisted methods by which these limitations are circumvented can nevertheless be transported to other machine learning tasks where problems of opacity, robustness, explainability, and missing data are critical. Moreover, given the transformative impact that causal modeling has had on the social and health sciences,[14,25,34] it is only natural to expect a similar transformation to sweep through machine learning technology once it is guided by provisional models of reality. I expect this symbiosis to yield systems that communicate with users in their

native language of cause and effect and, leveraging this capability, to become the dominant paradigm of next-generation AI.

Back to Top

Acknowledgments

This research was supported in part by grants from the Defense Advanced Research Projects Agency [#W911NF-16-057], National Science Foundation [#IIS-1302448, #IIS-1527490, and #IIS-1704932], and Office of Naval Research [#N00014-17-S-B001]. The article benefited substantially from comments by the anonymous reviewers and conversations with Adnan Darwiche of the University of California, Los Angeles.

Figure. Watch the author discuss this work in the exclusive *Communications* video.
https://cacm.acm.org/videos/the-seven-tools-of-causal-inference

Back to Top

Back to Top

Back to Top

References

1. Balke, A. and Pearl, J. Probabilistic evaluation of counterfactual queries. In *Proceedings of the 12th National Conference on Artificial Intelligence* (Seattle, WA, July 31-Aug. 4). MIT Press, Menlo Park, CA, 1994, 230–237.

2. Bareinboim, E. and Pearl, J. Causal inference by surrogate experiments: z-identifiability. In *Proceedings of the 28th Conference on Uncertainty in Artificial Intelligence*, N. de Freitas and K. Murphy, Eds. (Catalina Island, CA, Aug. 14–18). AUAI Press, Corvallis, OR, 2012, 113–120.

3. Bareinboim, E. and Pearl, J. Causal inference and the data-fusion problem. *Proceedings of the National Academy of Sciences 113*, 27 (2016), 7345–7352.

4. Chen, Z. and Liu, B. *Lifelong Machine Learning.* Morgan and Claypool Publishers, San Rafael, CA, 2016.

5. Darwiche, A. *Human-Level Intelligence or Animal-Like Abilities? Technical Report.* Department of Computer Science, University of California, Los Angeles, CA, 2017; https://arxiv.org/pdf/1707.04327.pdf

6. Graham, J. *Missing Data: Analysis and Design (Statistics for Social and Behavioral Sciences).* Springer, 2012.

7. Halpern, J.H. and Pearl, J. Causes and explanations: A structural-model approach: Part I: Causes. *British Journal of Philosophy of Science 56* (2005), 843–887.

8. Hutson, M. AI researchers allege that machine learning is alchemy. *Science* (May 3, 2018); https://www.sciencemag.org/news/2018/05/ai-researchers-allege-machine-learning-alchemy

9. Jaber, A., Zhang, J.J., and Bareinboim, E. Causal identification under Markov equivalence. In *Proceedings of the 34th Conference on Uncertainty in Artificial Intelligence*, A. Globerson and R. Silva, Eds. (Monterey, CA, Aug. 6–10). AUAI Press, Corvallis, OR, 2018, 978–987.

10. Lake, B.M., Salakhutdinov, R., and Tenenbaum, J.B. Human-level concept learning through probabilistic program induction. *Science 350*, 6266 (Dec. 2015), 1332–1338.

11. Marcus, G. *Deep Learning: A Critical Appraisal. Technical Report.* Departments of Psychology and Neural Science, New York University, New York, 2018; https://arxiv.org/pdf/1801.00631.pdf

12. Mohan, K. and Pearl, J. *Graphical Models for Processing Missing Data. Technical Report R-473.* Department of Computer Science, University of California, Los Angeles, CA, 2018; forthcoming, *Journal of American Statistical Association*; http://ftp.cs.ucla.edu/pub/stat_ser/r473.pdf

13. Mohan, K., Pearl, J., and Tian, J. Graphical models for inference with missing data. In *Advances in Neural Information Processing Systems 26*, C.J.C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger, Eds. Curran Associates, Inc., Red Hook, NY, 2013, 1277–1285; http://papers.nips.cc/paper/4899-graphical-models-for-inference-with-missing-data.pdf

14. Morgan, S.L. and Winship, C. *Counterfactuals and Causal Inference: Methods and Principles for Social Research (Analytical Methods for Social Research), Second Edition.* Cambridge University Press, New York, 2015.

15. Pearl, J. *Probabilistic Reasoning in Intelligent Systems.* Morgan Kaufmann, San Mateo, CA, 1988.

16. Pearl, J. Comment: Graphical models, causality, and intervention. *Statistical Science 8*, 3 (1993), 266–269.

17. Pearl, J. Causal diagrams for empirical research. *Biometrika 82*, 4 (Dec. 1995), 669–710.

18. Pearl, J. *Causality: Models, Reasoning, and Inference.* Cambridge University Press, New York, 2000; *Second Edition*, 2009.

19. Pearl, J. Direct and indirect effects. In *Proceedings of the 17th Conference on Uncertainty in Artificial Intelligence* (Seattle, WA, Aug. 2–5). Morgan Kaufmann, San Francisco, CA, 2001, 411–420.

20. Pearl, J. Causes of effects and effects of causes. *Journal of Sociological Methods and Research 44*, 1 (2015a), 149–164.

21. Pearl, J. Trygve Haavelmo and the emergence of causal calculus. *Econometric Theory 31*, 1 (2015b), 152–179; special issue on Haavelmo centennial

22. Pearl, J. and Bareinboim, E. External validity: From *do*-calculus to transportability across populations. *Statistical Science 29*, 4 (2014), 579–595.

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

24. Peters, J., Janzing, D. and Schölkopf, B. *Elements of Causal Inference: Foundations and Learning Algorithms.* MIT Press, Cambridge, MA, 2017.

25. Porta, M. The deconstruction of paradoxes in epidemiology. OUPblog, Oct. 17, 2014; https://blog.oup.com/2014/10/deconstruction-paradoxes-sociology-epidemiology/

26. Ribeiro, M.T., Singh, S., and Guestrin, C. Why should I trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (San Francisco, CA, Aug. 13–17). ACM Press, New York, 2016, 1135–1144.

27. Robins, J.M. and Greenland, S. Identifiability and exchangeability for direct and indirect effects. *Epidemiology 3*, 2 (Mar. 1992), 143–155.

28. Rosenbaum, P. and Rubin, D. The central role of propensity score in observational studies for causal effects. *Biometrika 70*, 1 (Apr. 1983), 41–55.

29. Shimizu, S., Hoyer, P.O., Hyvärinen, A., and Kerminen, A.J. A linear non-Gaussian acyclic model for causal discovery. *Journal of the Machine Learning Research 7* (Oct. 2006), 2003–2030.

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

31. Spirtes, P., Glymour, C.N., and Scheines, R. *Causation, Prediction, and Search, Second Edition.* MIT Press, Cambridge, MA, 2000.

32. Tian, J. and Pearl, J. A general identification condition for causal effects. In *Proceedings of the 18th National Conference on Artificial Intelligence* (Edmonton, AB, Canada, July 28-Aug. 1). AAAI Press/MIT Press, Menlo Park, CA, 2002,

567–573.

33. van der Laan, M.J. and Rose, S. *Targeted Learning: Causal Inference for Observational and Experimental Data.* Springer, New York, 2011.

34. VanderWeele, T.J. *Explanation in Causal Inference: Methods for Mediation and Interaction.* Oxford University Press, New York, 2015.

35. Zhang, J. and Bareinboim, E. Transfer learning in multi-armed bandits: A causal approach. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence* (Melbourne, Australia, Aug. 19–25). AAAI Press, Menlo Park, CA, 2017, 1340–1346.

Footnotes

a. Other terms used in connection with this layer include "model-free," "model-blind," "black-box," and "data-centric"; Darwiche[5] used "function-fitting," as it amounts to fitting data by a complex function defined by a neural network architecture.

b. One could be tempted to argue that deep learning is not merely "curve fitting" because it attempts to minimize "overfit," through, say, sample-splitting cross-validation, as opposed to maximizing "fit." Unfortunately, the theoretical barriers that separate the three layers in the hierarchy tell us the nature of our objective function does not matter. As long as our system optimizes some property of the observed data, however noble or sophisticated, while making no reference to the world outside the data, we are back to level-1 of the hierarchy, with all the limitations this level entails.

c. The assumptions encoded in Figure 3 are conveyed by its missing arrows. For example, Y does not influence X or Z, X does not influence Z, and, most important, Z is the only variable affecting both X and Y. That these assumptions lack testable implications can be concluded directly from the fact that the graph is complete; that is, there exists an edge connecting every pair of nodes.

d. Economists, for example, having chosen algebraic over graphical representations, are deprived of elementary testability-detecting features.[21]

About the Authors

Judea Pearl (judea@cs.ucla.edu) is a professor of computer science and statistics and director of the Cognitive Systems Laboratory at the University of California, Los Angeles, USA.