If you are interested in reviewing books for intelligence, please contact book review editor Karen Sutherland at intelligence_book-reviews@acm.org.

Causality: Models, Reasoning and Inference

Judea Pearl
Cambridge University Press, 2001
ISBN 0521773628
$39.95

Reviewed by: Joseph O'Rourke
Department of Computer Science
Smith College
Northampton, MA 01063
orourke@cs.smith.edu

Suppose you survey students in your class and discover that a higher proportion of students who smoke received a final grade of A than do students who do not smoke. Possible data are displayed in Table 1: 50 percent of the 10 smokers received an A, and only 40 percent of the five nonsmokers received an A. Puzzled by the seeming implication that smoking improves grades, you partition the same data differently, looking at students with high parental income (Table 3) separately from those with low parental income (Table 2). And you find even more surprisingly that the trend has been reversed: smoking lowers grades in both subpopulations. You have just encountered Simpson's Paradox: "an event C [smoking] increases the probability of E [grade A] in a given population p, and, at the same time, decreases the probability of E in every subpopulation of p" (p.174).

	A	≤ B	%A
S	5	5	50%
NS	2	3	40%

Table 1: A higher proportion (50 percent) of students who smoke (*S*) receive an A than do students who do nonsmokers (*NS*). Smoking improves grades?

	A	≤ B	%A
S	5	4	56%
NS	1	0	100%

Table 2: Low-income students: Smoking reduces the percentage of A's.

	A	≤ B	%A
S	0	1	0%
NS	1	3	25%

Table 3: High-income students: Smoking again reduces A's.

One of the great achievements of Judea Pearl's work is to dispel the cloud of mystery enveloping Simpson's Paradox for a century. He analyzes it so thoroughly that he even explains (p. 182) why we find the reversal in subpopulations paradoxical.

The engine driving Simpson's Paradox is *causality*, and the confusion derives from trying to understand it solely through statistics:

> *It is an embarrassing yet inescapable fact that probability theory, the official language of many empirical sciences, does not permit us to express sentences such as "Mud does not cause rain"; all we can say is that the two events are mutually correlated, or dependent—meaning that if we find one, we can expect to encounter the other* (p. 134).

We often analyze statistical data in search of causes: "Does smoking improve grades?" Pearl demonstrates conclusively that such data cannot be adequately analyzed through statements of the probability calculus such as

$$Prob(A \mid S) > Prob(A \mid NS)$$

Instead he proposes in this book a type of causality calculus, a precise mathematical language for expressing causal relationships and answering questions about them. The language is partly probabilistic, and partly graphical, employing *causal diagrams*, such as the diagram in Figure 1a, in which the directed

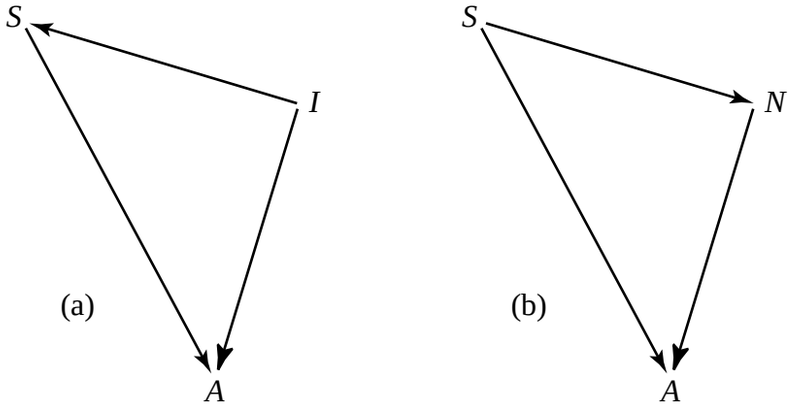


Figure 1: (a) *I* (Income) influences S (Smoking) and A (Grades); (b) S (Studying) influences N (Notetaking) and A (Grades).

links from Income to Smoking and Grades represent "direct functionally relationships" (p. 44) between the node variables.

Under this causal model, it is necessary to look at the income-separated data (Tables 3 and 2), because the effect of smoking upon grades is *confounded* (p. 182) by income.

However, suppose S represents Studying, and I is replaced by N (Note-taking behavior). A more plausible causal diagram might be that shown in Figure 1b: Studying affects note-taking and both affect grades. In this case the combined data of Table 1 should be used for analysis, for reasons detailed by Pearl.

Simpson's Paradox is just one of Pearl's success stories. "One of my main objectives in writing this book," he says, "is to see these confusions resolved" (p. 173), and he succeeds brilliantly. Let me provide two further examples.

Both social scientists and economists have developed languages for causality: the former, structural equation modeling (SEM); the latter, potential-outcomes models. Despite venerable 75-year histories, neither has caught on, inside or outside those fields: "the structural equation framework because it has been greatly misused and inadequately formalized, and the potential-outcome framework because it has been only partially formalized and... because it rests on an esoteric and seemingly metaphysical vocabulary of counterfactual variables..." (p. 134). Pearl proves that the two frameworks are mathematically equivalent (p. 243) and settles six outstanding questions about the interpretation of SEM (p. 170), which together delimit the bounds of the applicability of the approach. All of this is phrased within his calculus of causality, demonstrating its power and clarity.

The second impressive example is Pearl's clarification of the murky waters of *counterfactual* statements, such as: "the probability that event B would have been different if it were not for event A" (p. 27). Although such statements may seem abstruse, Pearl says that "it is worth emphasizing that the problems of computing counterfactual expectations is not an academic exercise; it represents in fact the typical case in almost every decision-making situation" (p. 217), a point given further support by the recent work of McCarthy (Costello and McCarthy 1999). Pearl's calculus of causality is especially well-suited to expressing counterfactuals, and he employs it to analyze the philosophical theory of causality developed by David Lewis (1973). Following a line present in the thought of David Hume and John Stuart Mill, Lewis proposed that we abandon attempts at using regularity to capture causality and instead interpret "A has caused B" as "B would not have occurred if it were not for A" (p. 238). Evaluating counterfactual statements under Lewis's theory involves comparing the similarity between various "possible worlds." His theory is notoriously intricate; it has generated a substantial body of commentary and criticism.

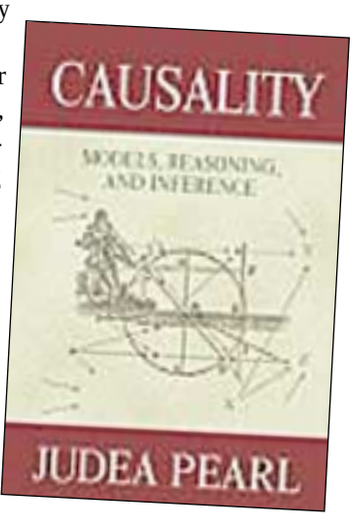
Pearl establishes the exact relationship between his own calculus and that of Lewis, showing that they are equivalent for "recursive models"; however, in nonrecursive systems,

Lewis's axioms do not imply an important "reversibility" property (p. 229). Further, he develops a precise meaning for the important notion *actual cause*—"the event recognized as responsible for the production of a given outcome" (p. 309)—and shows how to avoid failures of Lewis's "counterfactual dependence chains" to uncover the actual cause. Pearl identifies three primary types of questions that one might hope a causal theory would support (p. 29):

1. *Prediction* (Will a tornado form given particular weather conditions?)
2. *Intervention* (Will the economy rebound if the Federal Reserve lowers interest rates?)
3. *Counterfactuals* (Would we live longer if our salt intake were lower?)

These three tasks form a natural hierarchy on the causal model knowledge needed to address them. His book follows this hierarchy, starting with two chapters on prediction in causal Bayesian networks and ending with four chapters on counterfactual analysis. The heart of the book, in four chapters, concerns intervention, the study of which is greatly facilitated by Pearl's do operator. do(x = a) is an intervention that clamps variable x to value a. The semantics of Prob(y | do[x = a])—"the probability of y when x is set to a"—is quite different from the usual Prob(y | x = a)—"the probability of y given that we find that x is a." do( ) is an action, not an observation, an important distinction because "most scientific knowledge is organized around the operation of 'holding x fixed' rather than 'conditioning upon x'" (p. 98). He develops a formal do-calculus (p. 85), which, for example, specifies that Prob(y | do[x = a]) cannot be computed when there is a "back-door" confounding path in the causal diagram between x and y containing only unobserved variables. Such a clear encapsulation of a subtle issue again demonstrates the power of his framework.

This book could serve as the focus of a graduate seminar, but it would not easily support a more traditional lecture course. It is a research monograph, not a textbook. Many of the chapters are revisions of journal articles, with corresponding style and rigor. There are more than 50 lemmas, theorems, and corollaries, and nearly twice that number of technical definitions. Although it is literally true that "expert knowledge of logic and probability is nowhere assumed in this book" (p. xiv), this does not imply that the material is accessible to undergraduates or beginning graduate students. Even research professors will find it challenging but consistently enlightening.



I anticipate two directions for future development. First, Pearl's successes are largely theoretical, for example, proving that theory A is equivalent to theory B or resolving the confusions of theory C. The reader must wait 270 pages to reach the first uncontrived example. Pearl's approach must be field tested on a wide variety of real data sets to convert the skeptical. Now that the theoretical ground is cleared, substantive application examples will surely follow. Second, there is a need for expository versions of this story. The current price of admission into this fascinating world is rather high. No doubt insiders soon will write guides to make entry easier. But even then, I expect this book to stand as the primary authority on reasoning with causality for years to come.

References

Costello, T. and McCarthy, J. 1999. Useful counterfactuals. *Electronic Transactions on Artificial Intelligence 3*, 2.

Lewis, D. 1973. *Counterfactuals.* Harvard University Press.

* * * * * * * * *

Data Mining: Concepts and Techniques

Jiawei Han and Micheline Kamber
Morgan Kaufmann Publishers
San Francisco, CA, 2000
550 pp.
ISBN 1558604898
$54.95