Available online at www.sciencedirect.com
SCIENCE DIRECT®
Artificial Intelligence 169 (2005) 174–179
Artificial Intelligence
www.elsevier.com/locate/artint

Book review

Judea Pearl, Causality, Cambridge University Press, 2000.

Henry E. Kyburg Jr.

University of Rochester, Institute for Human and Machine Cognition, USA

Available online 28 October 2005

This is a remarkable volume. Winner of the Lakatos award, given biennially for the book in the philosophy of science most highly regarded by an international committee, it is also crammed with formulas that will be of practical importance, as well as of interest, to epidemiologists, lawyers, economists, and other down-to-earth folk. This is not to say that it will be easy to read, for anyone, or that it is altogether correct. I shall first offer a review of the contents of the book, and then carp (minimally) about the viewpoint. The review of the contents will be highly schematic, since the book is extremely rich.

Many of the results reported in the first six chapters here parallel results also achieved by researchers at CMU [4]. The first two chapters present background, and introduce the terminology of probabilistic networks. Chapter One introduces some probability, with a one page nod to subjectivism; conditional probability is glossed as "... given that I know A" [p. 5]. Subsequently the references are to the first person plural, which suggests a relatively objective conception of probability. Most often we encounter relations among probabilities (such as $P(y \mid x, z) = P(y \mid z)$) which make perfectly good sense as objective frequencies. The issue of subjectivism is one to which we will return later.

Among the basic ideas introduced in the first chapter are these: Directed Acyclic Graphs (DAG's); conditional independence; Bayesian networks; the Markov property, and d-separation. The very important operator $do(X = x)$ that introduces a way of treating *actions* is introduced on p. 23. This operator sets a subset X of variables to constants x, yielding an *interventional distribution.* Let P_* be the set of all interventional distributions $P_x = P(v \mid do(X = x))$ (including the empty intervention). A *DAG* G is a *causal Bayesian network* compatible with P_* if and only if the following three conditions hold for every P_x in P_*:

1. $P_x(v)$ is Markov relative to G;
2. $P_x(v_i) = 1$ for all $V_i \in X$ whenever v_i is consistent with $X = x$;

E-mail address: hkyburg@ihmc.us (H.E. Kyburg).

doi:10.1016/j.artint.2005.10.003

3. $P_x(v_i \mid pa_i) = P(v_i \mid pa_i)$ for all $V_i \notin X$ whenever pa_i (the set of parents of X_i) is consistent with $X = x$.

Given a *DAG* the simplest interventional distribution is produced by deleting the arrows into X, making the variables in X exogenous. The $do(\cdot)$ operator represents the most important idea in the book. It is used heavily in chapters three through six, concerned with identifying and computing the results of interventions, and is crucial in the final four chapters of the book, concerned with counterfactuals.

Section 1.4 introduces functional models, characterized by a set of structural equations: $x_i = f_i(pa_i, u_i)$, according to which the value of the quantity X_i is a deterministic function f_i of the values of the parents of X_i and the random (unmeasured) error u_i. Druzdel and Simon [1] (cited by Pearl, p. 31) showed that for every Bayesian network G characterized by a probability distribution P, there exists a functional model that generates a distribution identical to P. Pearl makes much use of this fact in the subsequent parts of the book.

The effect of the $do(X = x)$ operator on a set of functional equations is easy to represent: Merely delete the equations giving the values of the variables in X as a function of the values of their parents, replacing them with the values x.

The second chapter contains two important algorithms, *IC* and *IC**, that parallel algorithms presented by [4]. *IC* takes as input a stable ("faithful" in the terminology of [4]) distribution on a set of variables and produces a partially directed graph G. A stable distribution is one in which no independencies arise due to numerical coincidence. *IC** takes as input a stable distribution (with respect to some latent structure) and yields a marked pattern that is a core of this distribution. The pattern produced, in either case, singles out only an equivalence class of structures. These algorithms provide the basic connection between statistical data and causal graphs. Pearl puts it thus: "...we have shown that the assumption of model minimality, together with that of 'stability' (no accidental dependencies) lead to an effective algorithm for structuring candidate causal models capable of generating the data, transparent as well as latent" [3, p. 60]. Note that "the data" consist of statistical observations, not psychological assessments.

Chapters Three, Four, Five, and Six contain the computational heart of the book. Chapter Three, *Causal Diagrams and the Identification of Causal Effects*, is the meatiest chapter in the book. It provides a formal semantics for intervention (the $do(\cdot)$ operator) in terms of causal diagrams and probability distributions, and it gives formulas for postintervention probabilities in terms of preintervention probabilities. The effects of every intervention can be calculated if we have a DAG and none of the variables are latent. When some variables are latent the matter is not so simple; but conditions can be given for identifying causal relations. A bonus in Chapter Three is a calculus of interventions: a set of rules "...by which sentences involving interventions and observations can be transformed into other such sentences ..." [3, p. 65].

Chapter Four is primarily an extension and elaboration of some of the results of Chapter Three. It deals with actions, plans, and the distinction between direct and indirect effects of actions. Chapter Five is focused on structural models in Social Science and Economics. Causality has been regarded as problematic in the social sciences, despite the fact that one of the things we would like is that our empirical knowledge in these areas should inform our decisions—that is, that we should be able to anticipate the causal consequences of

our actions, such as raising taxes. The chapter is rich with historical asides, and Pearl's methods do throw light on a number of controversies in econometrics, epidemiology, and law. Section 5.4 approaches the question of the meaning of structural equations head on, and provides an explanation of the asymmetry of structural equations ($y = \beta x + \epsilon$ is not at all the same as $x = (y - \epsilon)/\beta$) and of the relevance of structural equations to policy making. Lacking background in either economics or the social sciences, I found this chapter hard going, and would have benefited from more, and more detailed, examples. Yet it was clear to me that the machinery introduced in chapters three and four had a great deal to offer the disciplines of the social sciences.

Chapter Six concerns the important topic of confounding: when is the apparent association between X and Y really due to (*caused by*) a third variable Z that influences both X and Y? The topic is introduced by a fascinating and thorough discussion of Simpson's "paradox", in which a drug proves beneficial to a mixed group of males and females, but deleterious to the males in the group and deleterious to the females in the group. As Pearl notes, there is nothing shocking about the existence of a probability distribution that represents the facts. What may be a bit unsettling is the fact that if we were to know only the general data, we would prescribe the drug. If we know the details, we would not prescribe the drug to women, and we would not prescribe the drug to men. What should we prescribe for Robin, a patient of unspecified sex? The answer in this case is that we should not prescribe the drug, but this reflects the causal structure we reasonably attribute to the process underlying the data. There are numerically similar cases in which a different causal structure leads to a different prescription. This chapter, as contrasted with the previous one, is rich in detailed examples that are readily accessible to the nonspecialist.

The last four chapters contain some of the most original and daring of the material in the book. They are concerned with "structure based counterfactuals". The general idea is that we can define the $do(\cdot)$ operator on causal models; that the effect of an action $do(X = x)$ is a submodel M_x in which X has been set to x; and that the effect on Y of that action is the value of Y in the submodel M_x; this is denoted by Y_x. In other words, "If X were x, then Y would be y" in model M holds just in case $Y = y$ is the solution of the equations characterizing M_x, the submodel reflecting the action $do(X = x)$.

After presenting an example (from econometrics), and some discussion of the value of counterfactual talk in explanation, Pearl presents (in 7.3) a set of axioms characterizing counterfactuals. These axioms are provably sound and complete for causal models. Of greater interest to philosophers (even fellow travelers) will be the comparison between structural counterfactuals and David Lewis's similarity based counterfactuals. Pearl shows that for recursive systems, Lewis's logic (as extracted from [2] by Pearl) satisfies the constraints imposed by Pearl's construction of counterfactual logic. Pearl's logic is narrower, since actions in structural models are limited to conjunctions of literals. Whether this is a loss is an open question. In terms of probabilities, actions, for Lewis, require the relation of imaging: given that an action rules out some worlds, the corresponding probability mass is transferred to the "closest" possible worlds, rather than being distributed among all the surviving worlds; but this is just what is done by updating a probability by the $do(\cdot)$ operator.

Thus Pearl's treatment of counterfactuals parallels much of what has been discussed in Philosophy. My own feeling is that while the differences may be of interest to those

philosophers who have found the sentential operator $\Box\!\!\rightarrow$ useful and its interpretation in possible worlds illuminating, most philosophers, and all scientists, will find Pearl's formalism far more accessible and natural. Surely from an intuitive standpoint it is natural to think of the consequent of a counterfactual as what will (would) be true when the antecedent is *forced* to be true. That is, causes seem, intuitively, to precede counterfactuals.

Less important than the main content of the book (but fun to read) are Pearl's pronouncements on Philosophy. For example, "Evidential decision theory was a passing episode in the philosophical literature" [3, p. 109]. But there are minor errors as well: nobody since Frege's attack on psychologism in logic has thought of deductive argument as argument "from beliefs to beliefs" [3, p. 209].

Chapter Eight primarily concerns the analysis of experiments in medical research. Chapter Nine returns us to the world of Philosophy and the treatment of necessary and sufficient conditions. Curiously, given how successfully thorough Pearl has been in his philosophical enquiries in the rest of the book, the formal work of G.H. von Wright [5] on necessary and sufficient conditions is overlooked. Of course Pearl is interested in the probabilities of necessary and sufficient conditions, which is not what von Wright was concerned with. A number of theorems are presented bearing on the identifiability of the probability of necessary and sufficient conditions, and helpful examples are discussed.

The last chapter, *The Actual Cause*, concerns singular causal claims: "Joe was killed in a car crash", as opposed to "Car crashes kill people". To make sense of this, it behooves us to think again of the nodes in the causal graph. The set U is the set of latent variables or quantities. They play the role of God; they are responsible for everything that happens. The $do(\cdot)$ operator changes the world, but in terms of the causal graph this fact is miraculous. The set V is the set of measured quantities. There is nothing bizarre about a measured quantity D that takes the value 1 if Joe is dead, and the value 0 otherwise. The use that Pearl makes of causal graphs depends on the assumption that there is a probability function defined on the nodes of the graph that is Markov (or semi-Markov). There is thus a perfectly well defined probability $P(D = 1)$, and a perfectly well defined conditional probability based on what we know, e: $P(D = 1 \mid e)$. The latter, of course, is 1, since we know that Joe is dead. What we need to do is to construct a model M_u—a "beam" in Pearl's terminology—in which $D = 1$ holds, given $do(Carcrash = 1)$, but $D \neq 1$ given $do(Carcrash \neq 1)$.

The recipe for constructing a "natural beam" in terms of which we can define *actual cause* in a given causal model $\langle U, V, \{f_i\}\rangle$ and state u, is this: for each variable V_i select a subset S_i of its parents, where S_i is sufficient to ensure the actual value of V_i regardless of the values of PA_i/S_i. We say that $X = x$ is an actual cause of $Y = y$ if and only if there is a natural beam M_u such that $Y_x = y$ (or $P(Y = y \mid do(X = x)) = 1$) and $Y_{x'} \neq y$ (or $P(Y \neq y \mid do(X = x')) = 1$) in M_u for $x \neq x'$. With a somewhat broader notion of causal beam, we can also define "contributory cause". This is all relative to a specific model u. If U_{xy} is the set of states in which "x is the actual cause of y" is true, and U_e is the set of states compatible with the evidence e, the probability that x is the actual cause of y is just the conditional probability $P(U_{xy} \mid U_e) = P(U_{xy} \cap U_e)/P(U_e)$.

The basic idea around which the whole book is built is that of a DAG, or, alternatively, a set of structural equations. These structures are tied to Bayes nets. As we have just seen, even in explicating causality we suppose that there is a single probability function P defined over the set of structures with which we are concerned. The machinery employed by

Pearl begins with directed graphs having probabilities defined on their nodes. It is in terms of such objects that causality is defined.

It is thus mildly disturbing to be told, on p. 25, that causal relations are more *stable* than probabilistic relations, "... because causal relationships are *ontological*, describing physical constraints in our world, whereas probabilistic relationships are *epistemic* reflecting what we know or believe about the world". While holding a subjective view of probability may incline one to talk this way, it rather undermines the force of being able to say something about the consequences of actions on the basis of mere beliefs. Does what we can say depend on "belief" in the sense of opinion? Are relative frequencies not just as "ontological" as causal relations? The relevant difference between causes and probabilities is that a causal relations hold always, and probabilistic relations hold only a certain fraction of the time. But there need be nothing "subjective" about that fraction; and that is exactly why the techniques provided by Pearl are objectively useful.

To be sure, "subjective" is ambiguous: it may mean produced by the mind or it may mean resulting from the feelings or temperament of the subject. In the former sense the objective measurement of length is subjective, in that at some point it involves a perceiving subject. In the latter sense it is contrasted with "objective" and is something we seek to purge from science. In the writings of "subjective Bayesians" subjectivity often seems to reflect a naive belief in *subjective idealism*—the extreme doctrine that nature has no objective existence independent of the perceiving mind. Pearl is much too smart and much too good a scientist to become trapped in the extreme view, but there are many instances in which he comes very close to endorsing the subjective Bayesian view.

"Traditionally, statisticians have approved of only one way of combining substantive knowledge with statistical data: the Bayesian method of assigning subjective priors to distributional parameters" [3, p. 96]. Before 1950 no respectable statistician would have approved of this, and even now it is probably safe to say that most statisticians would not approve of this approach. Nor should they.

"If ... the variance of X changes because we (or Nature) locally modify the *process* that generates X, then ... the ratio $\beta = E(YX)/E(X^2)$ will remain constant. However if the variance of X changes ... because we observed some evidence $Z = z$ that depends on both X and Y ... then that ratio will not remain constant" [3, p. 162]. In much of the volume probabilistic values represent relative frequencies. Relative frequencies do not change in response to our observations (though what we assert about them may).

"The conditioning operator in probability calculus stands for the evidential conditional 'given that we see,' whereas the $do(\cdot)$ operator was devised to represent the causal conditional 'given that we do' [3, p. 174]. Isn't it more plausible to take the conditioning operator to be representing submodels of a given model than to infuse probability theory with psychology? The $do(\cdot)$ operator modifies our model in a *different* way.

In introducing probabilistic causality, Pearl refers to "... the assumption that human experience is encoded in the form of a probability function ..." [3, p. 249]. True, he admits that this assumption is not entirely compelling, though as stated it is either patently false or a very deep theorem in the psychology of the unconscious.

Where it really counts, Pearl's thought, even in these dangerous waters, flies straight as an arrow. For example, discussing attempts to characterize causality subjectively, he writes "By far the most critical and least defensible paradigm underlying probabilistic causality

rests on the assumption that one is in possession of a probability function on all variables relevant to a given domain" [3, p. 252]. Again, the *definitions* that concern causal models [3, pp. 203–205] invoke no trace of the subjective point of view. Introducing the chapter on action, Pearl writes "In principle actions are not part of probability theory ... probabilities capture *normal relationships in the world*, whereas actions represent interventions that perturb those relationships" [3, p. 109, my italics]. This seems to me clear and precise and unexceptionable; it invokes no confusion between the psyche and the building blocks of the objective world. The "normal relationships" in the world might more clearly be captured in terms of relative frequencies. To apply the structures that Pearl has given us then would require the forging of a connection between relative frequencies and probabilities in the sense that probabilities are a guide to life, and that is not part of Pearl's program.

It would also be useful to discuss sets of models that would allow us to represent *approximate* relative frequencies. These, after all, are what we have good reason to accept. For example, we should take the relative frequency of heads on ordinary coin tosses to be *about* a half, rather than 0.5000.... We might represent the behavior of n tosses of the coin by a set of binomial models with a parameter *close* to 0.5. Note that although the frequency of HH in these models may be as low as $0.45 * 0.45$, the frequency of HT cannot be higher than $0.5 * 0.5$.

Am I carping? I am not just fishing for things to complain about; these are matters that seem to me very important in the application of Pearl's formalism. It is precisely because the models to which he introduces us reflect both the causal structure of the world and relative frequencies in the world that they are so useful. If the "probabilities" were merely subjective (personal, whimsical, temperamental) that would deprive Pearl's machinery of some of its most important uses. On the whole, this is a profound and enormously important book. It is not easy going, but it is perpetually rewarding. The subjectivism that occasionally rears its foggy spector does not seriously impede the clear-headed reader.

References

[1] M.J. Druzdel, H.A. Simon, Causality in Bayesian belief networks, in: Proceedings of the 9th Conference on Uncertainty in Artificial Intelligence, 1993, pp. 3–11.
[2] D. Lewis, Counterfactuals, Harvard University, Cambridge, 1973.
[3] J. Pearl, Causality, Cambridge University, New York, 2000.
[4] P. Spirtes, C. Glymore, R. Schines, Causation, Prediction, and Search, Springer-Verlag, New York, 1993.
[5] G.H. von Wright, A Treatise on Induction and Probability, Chapman and Hall, London, 1951.