Brit. J. Phil. Sci. **55** (2004), 779–790, axh413

work of Euclidean theories is confined to theorem-proving: axioms may be reformulated to separate and illuminate their contents or to provide justifications.

The results of Socratic theories, the final analyses, are almost never accurate and have no interesting consequences except their instances; their value, and it can be considerable, is in the sequence of arguments produced for them, the examples and counterexamples, the explorations of conceptual relations that justify or defeat them, and the approximate connections they articulate. The ubiquitous failure of philosophical analyses should be no surprise: they are confined almost exclusively to explicit, or occasionally simple inductive, definitions, and philosophically interesting concepts are seldom if ever strictly so related. Euclidean theories are almost always inaccurate as well, and that too should be no surprise, since universality without vacuity is rare. Yet, unlike Socratic theories, Euclidean theories have the virtue that their consequences can be interesting, informative and not obvious from their starting assumptions. Euclidean theories are justified in part by their fruits; Socratic theories have few, if any, except in the process that generates them. Socratic theories overwhelmingly predominate in contemporary analytic philosophy, but Euclidean theories occasionally arise and have some influence, as with theories of counterfactual inference, or axioms of subjective probability.

These divisions in philosophy are imperfectly correlated with a division in aims. Philosophy can aim to describe and clarify a conceptual pr /Fwobability.

chapters may give the impression that this is a standard Socratic work. It is not. The book begins that way, but then moves to a Euclidean mode. A second aim of the book is to apply pieces of the theory and Woodward's elaboration, specifically the ideas of intervention and invariance, to impose constraints on causal explanations, and to account for qualitative features of our assessments of the value of various explanations. These two goals are addressed in the book almost interactively, with discussions of key themes from the axiomatic theory mixed in with applications and illustrations of the proposed constraints and semi-Socratic explorations.

I will organize things a bit differently, first presenting the background to the axiomatic theory Woodward reconstructs; then briefly the theory itself in something close to its original form; next Woodward's analysis of its pieces and the general themes which he finds connected with those pieces, or which he thinks are needed to justify causal attributions, and finally his novel axiomatization and generalization of the theory's content. Following that, I will briefly describe some of the applications Woodward makes of component ideas and meta-themes for assessing causal explanations, and the constraints on scientific explanation he would impose. My only objections to the book are disagreement with some of his constraints on a causal explanation, and regret that it does not take account of more of the relevant statistical and computational literature.

2 Interventions and causal Bayes nets

In the 1980s, Donald Rubin ([1986]) developed a counterfactual analysis of causation and experimentation which has come to dominate statistics. It fit into the parameter estimation paradigm of statistics, but afforded no means of computing the effects of interventions in complex systems, or of computing the effects when unobserved causes also influence affected variables. That possibility was developed in the same period by Jaime Robins ([1986]), in work that allowed the computation of the effects on remote variables. Because of the difficulty of its presentation, Robins' work had little influence for some time. Neither approach offered much advance on how to search for and discover causal relations in the absence of experiments. Those developments depended on the formalization of Bayes networks.

Bayes networks are directed acyclic graphs (DAGs) whose nodes are random variables with a joint probability distribution subject to a restriction. That restriction, the Markov condition, limits the pairing of DAGs and probabilities: each variable is independent of its graphical non-descendants conditional on its graphical parents. Parts of the formalism are implicit in theories of linear relationships developed in many disciplines throughout the last century, for example in regression analysis, factor analysis, Sewell

Wright's path models, and elsewhere, although there is nothing in the Markov condition that limits it to linear systems. In the late 1970s, a group of statisticians which included T. Speed, S. Lauritzen and K. Vijayan among others ([1978]) developed the general formalism and named the Markov condition (which perhaps ought to have been called the Reichenbach condition.) Speed (Kiiveri and Speed [1982]) recognized that the linear causal models then popular in sociology were examples of such networks. In the late 1980s, Judea Pearl ([1988]) and his students described many properties of these networks, most importantly, an algorithm for deciding whether a DAG implies, by the Markov condition, any given conditional independence relation, and described data-driven search procedures for limited classes of graphs. Pearl explicitly denied that the graphs have an objective causal interpretation, on the grounds that he saw no way to distinguish associations of two variables produced by a causal relation between them from associations produced by an unobserved or unrecorded common cause. Pearl's reservation echoed a long philosophical tradition claiming that algorithms for scientific discovery are impossible, a claim argued explicitly by Hempel from the alleged impossibility of algorithms that correctly introduce novel unobserved variables or 'theoretical terms'.

A natural response to Pearl's worry was to connect the Bayes net representations with experimental interventions that could in principle distinguish causal connections and that would warrant a causal interpretation of the networks. That step was taken by Spirtes et al. ([1993]) who connected the Markov condition with interventions—essentially showing that an ideal intervention in a causal system could be represented as a particular kind of graphical and probabilistic relation that would allow computation of the probabilistic effects of interventions. Specifically, the theory of Causal Bayes nets requires two axioms and two definitions for understanding the predictions of a wide class of causal hypotheses, and a third axiom for their discovery. Versions of all three axioms were in Pearl ([1988]), but without the interpretation and theorems that make up the causal theory. The first two

of no other variables in G*, and extending the joint probability distribution Pr on G to a distribution Pr*, satisfying the Markov condition for G*, such that each value of Ix determines a unique value of X and, conditional on any value of Ix producing a value x of X, the conditional probability function $Pr^*(|X = x) = Pr(|X = x)$. An ideal intervention on X for G and Pr is a specification of a value for a policy variable for X. The third axiom, sufficient for discovery, is faithfulness: all of the conditional independence relations in Pr follow from the Markov condition applied to G.

Reformulating the theory of Bayes nets as a causal theory and introducing definitions would be idle without the demonstration that the reformulation and definitions permit the use of causal hypotheses in prediction and explanation and without showing that such hypotheses can be learned from data. These are among Woodward's own desiderata for a theory of causal explanation. To that end, Spirtes et al. (1) showed that the probabilistic effects of ideal interventions on a system represented by a DAG could be calculated from the Markov condition, in some cases even when the DAG contains unobserved variables, and that such interventions could in principle distinguish any causal hypotheses represented by directed acyclic graphs; (2) proved that with a natural probability measure on the parameters describing a probability distribution for a graph, the converse of the Markov condition—the faithfulness condition—holds with probability 1; and (3) proved that, assuming faithfulness, and given any family of probability distributions for which conditional independence relations can be decided, there is a feasible algorithm for recovering partial information about causal structure represented by a DAG from observed values of variables (for independent, identically distributed sample cases) even when, for all one knows beforehand, there are unobserved common causes at work. It is exactly the abstractness of the networks—the fact that they do not represent the degree of any causal influence by parameters—that makes them appropriate objects for automated search. Parameters, for example linear coefficients or conditional probabilities, can of course be attached to such networks, but that is best done after the skeletal causal structure is selected.

The result is an axiomatic philosophical theory relating causation, probability and interventions whose consequences are still an active area of research in statistics and computer science—and, yes, in philosophy. This work has been developed in many other ways by many others, with applications in genetics, biology, economics, educational research, sociology, space physics, psychology and elsewhere, and it has been supplemented with extensive and ingenious developments by Pearl and his collaborators, with applications to a range of philosophical issues. Hempel's argument was met in part by the demonstration of algorithms that correctly (in the large sample limit) discover the presence of some unobserved common causes, and other

algorithms that allow correct prediction of the effects of interventions when unobserved common causes are known to be present (Spirtes et al. [1993]). More recently, Hempel's argument has been further disabled by the discovery of algorithms that under specifiable assumptions correctly identify variables related only by unobserved common causes, introduce those unobserved variables explicitly, and identify their causal relations with one another (Silva et al. [2003]; Spirtes et al. [2000]).

The theory of causal Bayes nets is obviously not the whole, complete story about causal relations in science. Some graphical causal models are cyclic, not acyclic; some models have 'correlated errors' given no causal interpretation; many causal theories are in the form of differential or difference equations, etc. The theory implies no claim that one or another of these fundamental notions—cause, probability, intervention—is the place to start, no more than Euclid's axioms imply that line segments are more fundamental than circles. Woodward suggests that for understanding causal explanation, the notion of intervention is the place to start.

Woodward defines an intervention this way: I is an intervention on X with respect to Y if and only if

1. I is causally relevant to X

2. I is not causally relevant to Y through a route that excludes X

3. I is not correlated with any variable Z that is causally relevant to Y through a route that excludes X

4. I acts as a switch for other variables that are causally relevant to X. That is, for some values of I, X ceases to depend on the values of other variables that are causally relevant to X.

The primitives of the analysis are 'dependence', 'causal relevance', 'correlation' and 'route'. The last term can be decomposed in graphical representations, which is what Woodward has in mind: a route from A to B is a directed path from A to B in which each link signifies a direct causal relation. Clause 4 makes it clear that 'A is causally relevant to B' means that A is a cause of B. I, X, and Y are variable features of an individual system. Woodward has in mind a definite population of values of these variables, or some potential distribution of values, and presumably by 'correlation' he means something more general than correlation, a frequency dependence in an actual population, or absence of independence in a probability distribution.

There are several reasons to understand Woodward's definition, which, as he notes, is closely related to the notion of an 'ideal intervention' introduced by Spirtes et al. as a term of art rather than a Socratic conceptual analysis of an ordinary notion. Surely an intervention, in an ordinary sense, can change

the form of dependence of X on other causes of X—for example, change the conditional probability of values of X on values of other causes of X—without entirely disabling the other causes (this is allowed in Spirtes et al.). Surely, too, the intervention I can cause some intermediate I' which causes X—here the ambiguity of 'dependence' gets in the way of clarity. Finally, but not exhaustively, in an ordinary and scientific sense, an intervention might directly alter two or more variables.

With the notion of intervention in hand, Woodward defines 'direct cause' and 'total cause', this time following Pearl ([2000]). I paraphrase (to avoid some minor ambiguities) his proposal as follows:

(1) X is a direct cause of Y with respect to variable set **V** not containing X and Y if and only if there exist values **v** for variables in **V** and values x, x' for X, such that interventions that hold constant all variables in **V** at **v** and that vary X between x and x' change the value, or the probability of Y.

(2) X is a total cause of Y if and only if there is a possible intervention on X that changes the probability distribution of Y.

Since the definition of intervention presupposes the notion of 'route'—i.e., a sequence of direct causes—the definition of 'X is a direct cause of Y' appears circular. Woodward argues otherwise, on the grounds that the definition does not presuppose information about whether X is a direct cause of Y, only about other direct causes. Ok, the definition is ill-founded, not circular: it could never

variables—in other words, a causal statistical model—we cannot with the definitions/postulates produced so far predict anything about the probability distribution that would result from an intervention, nor can we predict the conditional probabilities after an intervention. We cannot, in other words, yet explain how such theories are tested experimentally or used in quantitative explanations, let alone how they could be learned from data that is in part or whole non-experimental.

So Woodward needs more, and he gets it. The Causal Markov condition would fill all of the lacunae above, and more, but rather than invoking it, Woodward introduces and discusses two key aspects: invariance and modularity. Only after clarifying those aspects does he introduce axioms representing them, axioms that imply the Causal Markov condition.

same causal hypotheses, at least not if causal relations are assumed to be modular, a property Woodward defines this way:

> A system of equations is modular if (i) each equation is level invariant under some range of interventions and (ii) for each equation there is a

by PM and PLI alone. Still more is needed for an understanding of how one can reason even with these simple theories.

Woodward supplements these principles with two others.

(PM2) When X and Y are distinct, Pr(X | Y and set (**Parents** (Y)) = Pr(X | set(Y) and set (Parents (Y))

(PM3) If X does not cause Y, then Pr(X | **Parents**(X) and set(Y)) = Pr(X | Y and **Parents**(X))

and observes that: 'we may think of the Causal Markov condition CM as the conjunction of PM and PM3' (p. 341). I assume he has in mind that the two postulates imply that:

(4) If X does not cause Y, then Pr(X | **Parents**(X)) = Pr(X | **Parents**(X) and Y)

which is the Causal Markov condition. PLI then follows by an obvious definition of the 'set' operation in terms of an ideal intervention, and PM2 is a logical consequence. The theory of Causal Bayes nets is recovered, but separated into novel pieces. Only now, with a full axiom set, can the theory explain elementary features of causal reasoning such as those noted above.

We gain something important from Woodward's development of the theory: an understanding of the interaction of the ideas of intervention and invariance in causal explanations; the idea of modularity in causal models in which the probability of each variable is a function of the values of its direct causes and a parameter for each such cause, and a characterization of an analogous idea in more general models. There is a good deal left out of the discussion, including the theory of search and discovery and generalizations of the theory that allow for cyclic graphs representing feedback systems, and generalizations that allow 'correlated errors' to which no causal interpretation is assigned. Other connections with the computer science and statistical literature are passed by—for example, the use of an idea of modularity close to Woodward's, although actually stronger, in Bayesian search procedures for causal Bayes nets (Heckerman et al.). Woodward does a considerable amount to invite philosophers out of their disciplinary cave into the sunlight, and perhaps it is churlish to complain that he did not do more. So much for the first aim I ascribe to the book.

give an analysis of the notion that a particular value of a variable in a causal system is actually a cause of the value of another variable. His discussion is built around ideas proposed by Pearl and by Hitchcock, but I find it clearer and simpler than their presentations. Separately, Woodward plausibly accounts for the fact that the period of a pendulum is explained by the law of the pendulum, the length of its arm, and the gravitational field, but the period cannot, with the law and the gravitational field, explain the length of the pendulum arm: interventions that change the length of the pendulum arm change the period, but there are no interventions that change the length of the pendulum arm by intervening on the period. A similar analysis applies to

a cause, he argues, are typically ambiguous. Granted, but philosophers are skilled at disambiguating when they want to, and I think the point remains that genotype is not, on his view, even a remote cause of an individual's treatment by others. These last cases are in my view regrettable consequences of trying to found a theory of causal explanations on interventions. The alternative, perhaps not the only one, but a good one, is to regard 'X is a direct cause of Y with respect to variables V' as an unanalyzed primitive relation, to subject concatenated causes to the Markov condition, and to define 'X is an intervention with respect to Y in system G' as a particular kind of direct cause respecting the Markov condition. The result is an axiomatic theory in which race and sex and mineral composition can be causes—indeed, the very theory developed by Spirtes et al.

There are philosophers—Colin McGinn and Ronald Giere come to mind—who have claimed that only Socratic theories are philosophy. Woodward's