CLARK GLYMOUR

RABBIT HUNTING

Twenty years ago, Nancy Cartwright wrote a perceptive essay in which she clearly distinguished causal relations from associations, introduced philosophers to Simpson's paradox, articulated the difficulties for reductive probabilistic analyses of causation that flow from these observations, and connected causal relations with strategies of action (Cartwright 1979). Five years later, without appreciating her essay, I and my (then) students began to develop formal representations of causal and probabilistic relations, which, subsequently informed by the work of computer scientists and statisticians, led eventually to a practical theory of causal inference and prediction, a theory incorporating some of the sensibilities Cartwright had voiced (Glymour et al. 1987; Spirtes et al. 1993). That theory, and ideas related to it, have become a subfield of computer science with contributions far deeper than mine from many sources, and its inferential and predictive techniques have been successfully applied in biology, economics, educational research, geology and space physics.

My timing was bad. Sometime in the interim, Cartwright abandoned her earlier views. It is only natural that when a philosopher abandons an opinion, she should criticize the views of those who retain or extend it, and perhaps for that reason in recent years Cartwright has made the work of my collaborators and myself the object of repeated lengthy (but invariably courteous) criticism, first in a chapter of her second book, (most of which, she wrote, was intended to explain the "philosophical reasons" why the ideas behind our initial methods could not work), then in several papers, and soon in a chapter of a forthcoming book. With the exception of a short comment on one of her examples in *Nature's Capacities and Their Measurement*, I have written nothing in response. But a decade of criticism from a source so eminent eventually demands a considered reply. I will attend to the book just mentioned, and to a chapter of the manuscript of her new book, which contains the essential content of the intervening papers.

Synthese **121:** 55–78, 1999.

1. DISCOVERING CAUSAL STRUCTURE (1987) VERSUS NATURE'S CAPACITIES AND THEIR MEASUREMENT (1989)

An idea almost exactly as old as the twentieth century is that, in some circumstances, correlations among measured features may be explained by postulating that each measured variable is a linear function of unobserved causes that influence two or more measured variables, and of independently distributed unobserved causes particular to that measured variable. The joint probability density on all of the variables is determined by the joint density of the unmeasured variables and by the linear coefficients. The idea is basic to factor analysis, and is still a common model in psychometrics and elsewhere in social science. Later generalizations allowed that some of the measured variables might also influence others.

In such theories, neither the joint probability density, nor the causal relations between variables, nor the coefficients in the linear equations, are observed. The immediate scientific issue was how aspects of the unobserved structure might be estimated from observed associations among the measured variables. Charles Spearman, who introduced the representation, had the following idea: When four (or more) variables have a common cause, and there are no other common causes of any two (or more) of the variables, and the measured variables do not influence one another, then the model implies three equations among the correlations of the four measured variables:

$$r_{ij}r_{kl} = r_{ik}r_{jl} = r_{il}r_{jk}.$$

The implication holds no matter what values the linear coefficients may have, and no matter what the joint distribution of latent variables may be, so long as the correlations remain defined. One could, Spearman thought, confirm or disconfirm the causal explanation by testing these "tetrad equations" (as he called them) on sample data. His tests always assumed that the measured variables are jointly normally distributed, an assumption whose reasonableness could be judged by examining the data. Once the causal structure was established, the linear coefficients could be estimated from the data.

Spearman allowed only models with a single common cause. To accommodate subsequent empirical findings, his followers allowed models in which two or more uncorrelated, unobserved, common causes influence some of the members of the same foursomes of measured variables. The method died after Thurstone introduced factor analysis, largely because it was computationally intractable when, as was typical in psychometrics, a large number of measured variables were involved. In the 1960s, Herbert Costner realized that models with other unobserved causal structures

may similarly imply characteristic sets of tetrad equations. For example, a model in which two measured variables, i and j, share a common unobserved cause, and two other measured variables, k and l, share a second unobserved cause, implies a single tetrad equation:

$$r_{ik}r_{jl} = r_{il}r_{jk}$$

no matter how the unobserved common causes are causally related to one another. Costner and Schoenberg used this observation to propose heuristic methods for modifying initial latent variable models that fail statistical tests on sample data.

The general methodological point of *Discovering Causal Structure* was that social scientific practice is unnecessarily dogmatic, that usually very few alternative explanations are entertained, and that there are systematic, algorithmic ways to find, from among a vast logical space of possible causal models, those alternative explanations that can account for the data. One major technical contribution provided a feasible general algorithm for computing the set of tetrad equations implied by any (recursive) linear model in which unobserved causes (noises, or errors in other terminology) that directly influence only a single measured variable are distributed independently of one another. Despite the long history of efforts by Spearman and his followers, and the stimulus of Costner's work and prestige, no such procedure existed before. Our result was subsequently strengthened by several writers. The other major technical contribution of the book was to use these calculations, in combination with a sequence of statistical tests of the implied tetrad equations, in a data driven heuristic search procedure for modifying an initial latent variable model.

The heuristic search we proposed was essentially rather simple. It was easy to prove that adding a new functional dependency – a new causal connection – to an existing model may reduce the set of tetrad equations implied, but will never increase them. Starting with an initial latent variables model M, compute the set of tetrad equations among measured variables implied by M. and test them. Let the set of tetrad equations implied by M be I, and let the set of tetrad equations in I that survive the testing be H. Generally $I \supseteq H$. Find all of the models M' that extend M by adding additional dependencies and that imply only maximal subsets of H. The theoretical effort in designing this search lay in finding feasible ways to compute all such extensions of any given model M. In this, we were greatly aided by representing the functional dependencies of linear models by directed graphs.

There was one rather minor further technical contribution. Following ideas proposed by Herbert Simon, around 1960 Hubert Blalock introduced the idea of searching for linear models by testing the vanishing

partial correlations each model implies. Blalock's procedure was similar to Spearman's and Costner's in that it did not depend on the particular values of the linear coefficients. Blalock carried out his procedure only for simple models with no more than four variables, and he and subsequent social statisticians provided no general algorithm. We provided a general algorithm for computing the first order vanishing partial correlations implied by any recursive linear model without unobserved common causes and with independent errors, a result that seems pitiful in retrospect. Like Blalock, we offered no algorithmic procedure for using these constraints in searching for causal explanations.

The rest of *Discovering Causal Structure* was devoted to justifying heuristic search, explaining the procedures and the methodological intuitions behind them, illustrating their application on well studied sets of social data, and giving proofs. In applications, the search procedures were used to find models with free parameters (the linear coefficients and the variances of unobserved variables, assuming the normal family of distributions). The numerical values for the parameters were then estimated, and the fit models estimated, using a standard statistical package. The illustrative applications typically found plausible, better fitting alternatives to causal models in the social science literature. The single empirically independently verified application of the method was to predict, without prior knowledge, the order in which several questions in a famous sociometric questionnaire had been asked. Perhaps we were lucky.

Nancy Cartwright devoted sixteen pages of *Nature's Capacities and Their Measurement* to criticizing *Discovering Causal Structure.*

> The first and most important difference between my point of view and that argued in *Discovering Causal Structure* has already been registered. I insist that scientific hypotheses be tested. Glymour, Schemes, Kelly and Spirtes despair of ever having enough knowledge to execute a reliable test. (1989, 72)

What Cartwright described as our ill-founded "despair" was our emphasis on this: the fact that a statistical model passes a significance test at some alpha left is insufficient for the truth of the model, since many distinct models may pass the same test, and conventional statistical methodology had no method of finding the alternatives. That is true, and Cartwright said nothing to rebut it.

Next, simplicity.

> They assume that structures that are simple are more likely to be true than ones that are complex. I maintain just the opposite . . . have argued that nature is complex through and through: even at the level of fundamental theory, simplicity is gained only at the cost of misrepresentation . . . Glymour, Schemes, Kelly and Spirtes believe that simpler models are better. But I agree with Haavelmo. Simplicity is an artifact of too narrow a focus. (1989, 72–3)

We used the idea of simplicity exactly as follows: when a set of elaborations M_1 an initial model implies a maximal subset of H, and elaborating M_1 with further directed edges or functional connections yields a model M_2 that implies the same maximal subset of H, our procedures report M_1 but not M_2. One reason was purely computational. The output of the procedures would become unreadably large otherwise. But another reason was our belief that, absent substantive assumptions requiring particular connections, researchers would usually have no interest in such models. They would typically be unidentifiable, and be viewed as gratuitously complex.

Cartwright is perhaps correct that the whole truth about anything is very complex; but, quite properly, science is seldom interested in the whole truth, or aided by insistence upon it. In my view, an inquiry that correctly found the causes of most of the variation in a social phenomenon and neglected small causes would be a triumph; in her view it would be a debacle. In my view, anti-Newtonians who objected that there must also be magnetic forces on the planets, or that Newtonian theory does not explain the variations in the colors of planets, or their masses, would have missed the point; in her view, it seems, they would be making it. Still, she may have been correct that it would be better if social scientists had to consider more complex models as well as simple alternative models. There is no good reason why all parameters in the true, or even any nearly true, model of social or other phenomena must be identifiable. But it seems a bit churlish to have made that complaint of an effort that considerably expanded the ability to find alternative models, while making no comment on conventional practice.

Cartwright did not advance a more telling objection to our use of simplicity: a more complex model that implies the same tetrad equations as a simpler model might imply a distinguishing set of higher order, nonquadratic, constraints on the correlations, and our procedures did not test for such constraints, because we did not know how to compute them or test them.

Cartwright mixed her criticism of simplicity with criticism of a related methodological idea, which we called Spearman's principle, and which we put this way: "Other things being equal, prefer those models that, for all values of their free parameters (the linear coefficients), entail the constraints judged to hold in the population" (Glymour et al. 1987). In the models we considered, the principle is equivalent to ignoring or rejecting explanations of vanishing correlations, or of tetrad equations or vanishing partial correlations, that posit multiple causal pathways whose several influences perfectly cancel one another. She had a number of objections to

the principle, some of which she illustrated with a discussion of a case we considered, the Transitional Aid Research Project (TARP).

Separately in Texas and in Georgia, randomly selected groups of newly released fellows were given monthly payments for six months through the respective state unemployment commissions. After a year, rearrest rates for these groups were compared with rearrest rates for felons released at the same time in the respective states. In Texas there was no difference in rearrest rates between the two groups, and, likewise, in Georgia there was no difference in rearrest rates for the two groups. No data were obtained on the actual employment in this period of either the treatment or the control groups. The project leaders concluded, nonetheless, that these facts showed that payments to newly released felons reduce crime. They justified that odd conclusion in this way: in the experimental set-up, payments through the unemployment commission reduced the recipients propensity to work (supposition); unemployment caused the recipients to engage in crime (sociological theory); but since there was no difference in recidivism between the groups that received payments and the groups that did not (empirical data), the payments must have caused a compensating tendency not to do crime (conclusion). The two mechanisms perfectly canceled one another. The explanation is a straightforward violation of Spearman's principle.

In protest to these inferences Hans Zeisel, an eminent sociologist, very publicly resigned from the committee overseeing the experiment. Zeisel thought the straightforward and obvious explanation of the data was that payments (at least at the amounts in the experiments) do not influence recidivism. Our methods agree with Zeisel's in rejecting the arguments and the conclusion of the project leaders, and in thinking the experiment is evidence that payments have no influence on recidivism, but we went on for two pages to dispute Zeisel's claim that randomized experiments always have univocal interpretations, and that the only possible interpretation of the outcome of the TARP experiment is that there is no effect.

Here is Cartwright's assessment of our discussion:

> If unemployment really does cause recidivism, then, given the lack of correlation, cash-in-pocket must inhibit it; and if unemployment does not cause recidivism, then cash-in-pocket is irrelevant as well. The statistics cannot be put to work without knowing what the facts are about the influence of unemployment; and there is no way to know short of looking. Glymour, Schemes, Kelly, and Spirtes advocate a short cut. For them, it is more likely that unemployment does not cause recidivism than that it does. That is in part because of their 'initial bias against causal connections'. But the hypothesis that unemployment does not cause recidivism is as much an empirical hypothesis as the contrary; and it should not be accepted, one way or the other, unless it has been reliably tested. Failing such evidence, how should one answer the question, 'Does cash-in-hand-inhibit recidivism?' Glymour, Scheines, Kelly and Spirtes are willing to claim, 'Probably not' (1989).

Consider then the probabilities of hypotheses. Although our book did not give a Bayesian analysis, it is straightforward to do so. The probability density over a model with numerical coefficients (rather than free parameters) can be factored into a density for the model with free parameters and a conditional density for the parameter values. The methods we used correspond to constraints on prior probabilities over models with numerical coefficients (rather than free parameters), priors that put zero probability on models whose coefficients represent canceling mechanisms. In so far as the data force the posterior distribution of the linear coefficients (for the confounded model with free parameters, the model in which payments cause unemployment which causes recidivism) to be located close to a set of perfectly canceling parameter values, the posterior of the confounded model with free parameters approaches zero. With those priors, and the data of the experiment, "probably not" is the right answer. Put Bayesianly, Cartwright announced she had different priors, but it is difficult to see how inquiry could be conducted with them. Inferences from observational studies would be hopeless, since unobserved. causes can always be postulated that perfectly cancel observed associations. Potential confounding similar to the TARP experiment occurs in any randomized controlled experiment in which there is anything different about the treatment and control groups other than the intended treatment itself, and in social experiments there almost always is some difference. On her view of the complexity of things, we should expect such confounding to be typical, even if we are in ignorance of the confounding properties, in which case we will never be able to conclude from a randomized experiment that the treatment has no influence on the outcome.

Cartwright had a broader objection to our methods. She said we introduced a new "theory form".

> Here is how I would describe what they do. To get the new theory form, start with the old linear equations but replace all the usual continuous valued parameters in the equations by parameters that take only two values, zero and one. One can think of these new parameters as boxes, where the boxes are to be filled in which either a yes or a no; yes if the corresponding causal connection obtains and no if it does not. A specific theory consists in a determination of which boxes contain yes and which no (1989, 76).

That is indeed almost the form of the features of theories over which we search. But it is hardly new; in fact it is standard. Any reading of the social science literature shows the most common forms of linear models postulate equationals with linear parameters ranging over real numbers, not real numbers as linear coefficients, and that any number of possible linear parameters do not appear at all, because the model specifies that their values are each zero. (The parameter values are typically estimated

after the model with free parameters is specified.) With independent errors, that representation is *isomorphic* to the directed graphical representation we use. And, of course, the graphical representation is not new either; it was sixty years old when we used it and perfectly common, just as we represented it, in the social science literature in the thirty years preceding our book.

So what is wrong with our "theory form"?

> The upshot of this implementation of Spearman's Principle is to reduce the information given in a causal theory from that implied by the full set of equations to just what is available from the corresponding causal pictures . . . This move from the old theory form to the new one is total and irreversible in the Glymour, Schemes, Kelly and Spirtes methodology, since the computer program they designed to rank causal theories chooses only among causal structures. It never looks at sets of equations, where numerical values need to be filled in. I think this is a mistake, both for tactical and philosophical reasons (1989, 76).

There is nothing "total and irreversible" about the graphical representation that severs it from equational representations with free parameters. Four pages (68–72) of *Discovering Causal Structure* are devoted to describing how to transform graphical models into statistical models with equationals. We preferred the graphical framework because formal relationships are then easier to see, because directed graphs are natural for algorithmic work, and because no matter how the structures are represented, the proofs of essential properties are essentially graph theoretic. And there is, therefore, also no "total and irreversible" disconnection of the graphical representation from causal models with real coefficients rather than free parameters. They are what you get when you estimate the parameters in the model described by a graph. Absent only the explicit graphical representation, that is how linear models with a causal interpretation have been constructed almost everywhere and almost always in the social sciences, and that is why the theory of point estimation has played so important a role. Cartwright implies idiosyncrasy where there is nothing more than mathematical explicitness. All that is new is the methods of search, which use only the statistical features that are captured by the graphical (or equivalently, the equational structure), and the assumption that exogenous variables and noises are jointly independent. If you locate rabbits by their ears, you aren't implying they don't have tails.

Cartwright goes on briefly to explain, astonishingly, that correcting this "mistake" is the main point of most of her book:

> The philosophical reasons are the main theme of the remaining chapters of this book. The decision taken by Glymour, Schemes, Scheines, Kelly and Spirtes commits them to an unexpected view of causality. It makes sense to look exclusively at causal structures (i.e, their graphs) only if one assumes that (at least for the most part) any theory that implies the data from the causal structure alone is more likely to be true than one that uses the numbers

as well. This makes causal laws fundamentally qualitative: it supposes that in nature only facts about what causes what are important; facts about strength of influences are set by nature at best as an afterthought. I take it, by contrast, that the numbers matter, and that they can be relied on just as much as the presence or absence of the causal relations themselves (1989, 76–7)

Parallel reasoning to Cartwright's: It makes sense to try to catch rabbits by their ears rather than their tails only if they are more likely to have ears than tails. Our reasons for search over graphical structures rather than systems of equations had nothing to do with whether the existence of causal relations is more real than numerical measures of their strength, whatever that means; the reasons had everything to do with reliability and computational feasibility of search.

The value of searching over graphical structure can perhaps be illustrated by considering the numerically based algorithmic model elaboration procedures standardly used in 1987 (and even now), when *Discovering Causal Structure* was published. Cartwright did not mention them. Their strategy begins with a linear, latent variable model with free parameters and tests the model on sample data at some specified alpha level. In 1987, such tests required a computationally intensive numerical procedure for estimating values of the free parameters in the model. Implicitly, such models also contain many fixed parameters, usually corresponding to linear parameters whose values the model claims are zero, representing possible causal connections whose existence the model denies. If the model fails the hypothesis test, one of these previously fixed–at–zero parameters is freed. The parameter selected is whichever results in an elaborated model that most improves a fitting statistic. If two parameters are tied by this measure, one is chosen arbitrarily. The procedure iterates until no additional parameter results in a model against which the previous model is rejected by a hypothesis test. At each stage the computationally intensive numerical estimation procedure must be repeated. Because of the computational requirements, the search never branches: if freeing any one of several parameters results in equal improvement in search, only one is chosen and the others ignored at that stage of the search. And for the same reason, the procedure never has a model stage with more than one parameter free than in the models of the previous stage, even though such models will sometimes fit better than any model obtained by twice in succession freeing the best fitting single parameter. The result is a one step look ahead beam search.

In simplest form, our 1987 procedures began with the graph of the same initial latent variable model, tested the tetrad equations it implied, and stored the set I of those implied by the initial model and the subset H of I of implied tetrad equations that passed the test. For each single

directed edge that, when added to the initial graph, implied all members of H and a proper subset of I, a model M' containing that edge was created, and its corresponding set I' stored. The procedure then branched over all the elaborated models and repeated the process. A branch of search terminated when no further elaboration could reduce the implied tetrad equations without reducing the implied tetrad equations in H.

The critical difference in the procedures is that, for computational reasons, the conventional search could not afford to branch, and so had to make arbitrary choices, whereas our search, which required no numerical analysis, could and did branch. The difference in search strategies made a considerable difference in reliability. In an enormous simulation study using structures typical in social science models, with randomly assigned parameter values and a variety of sample sizes, the popular beam searches produced the correct answer in 11 to 13% of cases, depending on sample size and the particular algorithm used. Graphical search produced a set of alternative models containing the true model in more than 90% of the cases. Even if a single one of the alternative models output by our procedure were chosen at random for comparison with the single output of the beam search, the graphical search was correct in more than 40% of the cases.

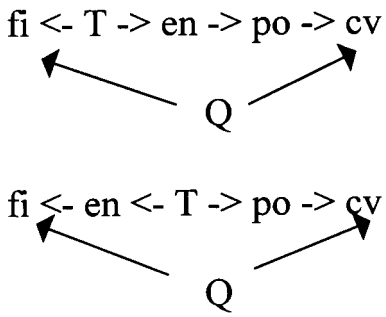
The essential issue in scientific discovery is the right representation for reliable, efficient search, not the metaphysical disputes upon which philosophy of science is fixated. Its Branching, not Being.

Cartwright almost concluded her criticism in *Nature's Capacities and Their Measurement* with a discussion of our treatment of another case, a study by Timberlake and Williams, purporting to show that, in the 1970s, foreign investment in "peripheral" nations caused "political exclusion" within those countries. The entire point of our discussion of the case, emphasized again and again, was that regression models assume a lot about causal structure, that if those assumptions are wrong, the regression coefficients do not measure causal influence, and that our methods sometimes permit one to find alternative explanations of the data, arguably sometimes better explanations. Cartwright claimed we used the case to criticize "methods, like those I have been defending, that try to infer causes from partial correlation" (1989, 79). The implication, which perhaps she did not intend, is that Timberlake and Williams used the methods of which she approved. I won't hold her to it.

Timberlake and Williams constructed measures of four variables, political exclusion (po), foreign investment (fi), energy development (en), and (absence of) civil liberties (cv), as well as other variables that do not figure in the argument. (They do not say which 72 nations they used, and we

could not reconstruct their data from their source. They did not reply to a request from my collaborator, Richard Scheines, for details on how the correlations were obtained.) They simultaneously linearly regressed po on fi, en, and cv, found a positive regression coefficient for fi, and concluded that they had shown that foreign investment caused political exclusion.

Examining their correlations, we found that political exclusion and foreign investment are uncorrelated when energy is controlled for, and that energy and absence of civil liberties are uncorrelated when political exclusion is controlled for. We said these vanishing partial correlations, which are very robust, "are the kind of relationship among correlations that can be explained by causal structure" (1987, 177), we offered some models that explain them in that way. For example, graphically:



where *T* and *Q* are unobserved causes. We did not claim any of these models are true, but did claim they are better explanations of the correlations and time order constraints reported by Timberlake and Williams than is the regression model in which the causal structure is assumed a priori and the vanishing partial correlations are accommodated by the numerical values of the coefficients. Since it involved no automated search, our analysis was exactly the sort that Hubert Blalock could have given.

The logic of Cartwright's discussion is difficult to follow. There is a formal point that may have been what she was after, namely that there exist (normal) probability distributions that do not satisfy Spearman's principle for any directed acyclic graph, with or without latent variables.

I will give her discussion in the sequence she did, changing only notation to agree with mine. First Cartwright asked the reader to assume, contrary to fact, that foreign investment and political exclusion are uncorrelated. Then she asked that the reader assume that two causal claims of the regression model are correct – en is a direct cause of po and cv is a direct cause of po, although there is nothing but the regression model to justify these assumptions. Then she asked the reader to assume that the

following three second order partial correlations do not vanish, although she made no showing of this assumption from the data:

> fi, po controlling for en and cv
> en, po, controlling for fi and cv
> cv, po controlling for en and fi

Each of their [i.e., our] structures reverses the causal order of (po) and (cv) (from the order in the TW model) . . . Since the methods described in Chapter 1 (of *Nature's Capacities and Their Measurement*) assume that temporal order between causes and effects is fixed, a structure in which (fi, cv and en all precede po), as they do in (Timberlake and Williams' model), will serve better for comparing the two approaches (1989).

Her point is that in our models po causes cv, whereas in Timberlake and Williams' regression model the reverse is true. She did not note that they gave no basis for their assumption.

There follows in her book a new graphical model which, she says, implies that po and fi are uncorrelated and also implies the two vanishing partial correlations we found from Timberlake and Williams correlation matrix. The model is:

$$\text{en po} \leftarrow T \rightarrow \text{cv} \leftarrow \text{fi}$$

She did not note that, unlike the time order of po and cv about which Timberlake and Williams provided an assumption but no information, they *did* specify that fi is measured at a later time than cv. Unlike our models, Cartwright's model really does violate what is known about the time order.

This structure, she wrote, implies the vanishing correlation of fi and po (which she assumed contrary to fact) and the two vanishing partial correlations (which we found in Timberlake and Williams' data) "just from its causal relations alone, keeps the original time-ordering and builds in the hypotheses favored by Glymour, Schemes, Kelly and Spirtes that investment does not produce repression" (1989, 82). That is true. So?

The example is an unfortunate one for Glymour, Schemes, Kelly and Spirtes, however. For there is no way that this graph can account for the (three non-vanishing second order partial correlations she assumed) with or without numbers. Nor is it possible with any other graph, so long as the time precedence of (en, fi and cv over po) is maintained. If the original time order is not to be violated any model which accounts for (the vanishing first order partial correlations) on the basis of its structure alone, and is consistent with (the nonvanishing second order partial correlations) as well, must include the hypothesis that foreign investment causes (1989, 82).

The obscure point is that in every linear model that implies

$$\rho \text{ fi,po} = 0$$
$$\rho \text{ po,fi.en} = 0$$
$$\rho \text{ cv,en.po} = 0$$

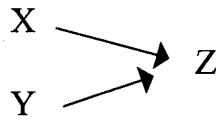
and does not imply

ρ fi,po.en,cv = 0
ρ en,po.fi,cv = 0
ρ cv,po.en,fi = 0

and does not have po cause fi, en or cv,

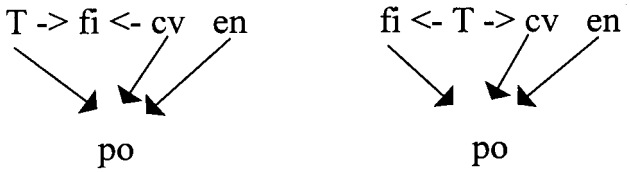
ρ fi causes po.

The assumptions are overkill. In any linear model, a structure of the form



implies that X, Y are correlated controlling for Z. (It does not matter whether the associations between X, Y, and Z are produced by X, Y causing Z or by a unmeasured common causes, or both.) Hence no linear model in which cv and en are each correlated with po, and po is not a cause of either of them, can imply that cv, en.po = 0. But why should it matter to our proposals that a completely imaginary set of constraints should not be explicable by any linear model? Presumably it should matter only if such circumstances are common, and there is some other method for finding the true structure when they arise. Cartwright did nothing either through an empirical survey or through mathematical analysis, to show that such constraints commonly occur. (Six years later we proved that, in the natural measure on the linear coefficients, such constraints have probability zero. See below.)

She continued by offering still another model which she claimed includes our "favoured hypothesis, that foreign investment does not cause repression, and does account for all the data, though of course not on the basis of structure alone" (1989, 83). The point seems to be a charitable effort on her part to formulate a model that saves the data (although it is not clear which data, real or imaginary, she meant to save), incorporates our "favored hypothesis" (although we had no stake in the particular causal claim, only a preference for certain explanatory relations), and corresponds to a time order which she seems to have thought was independently known. She presented the model graphically, as on the left below



and claimed that the data (whichever) cannot distinguish between this structure and Timberlake and Williams' model. Her version of Timberlake and Williams' model – which is not the real one – is shown on the right above. Her version leaves out the correlations of fi and cv with en implicit in the regression model.[1] And what follows?

Cartwright might have been trying to show that there are probability distributions that cannot be explained by any directed, acyclic graphical model consistent with Spearman's principle. There are. She might have been trying to show that there are probability distributions that: cannot be explained by any graphical model consistent with both Spearman's principle and a substantive assumption. There are. The relevance of this discussion to our methods is mysterious, absent a showing that such patterns of constraints are typical.

2. CAUSATION, PREDICTION AND SEARCH VERSUS MORE RECENT CARTWRIGHT

In several places throughout her discussion in *Nature's Capacities and Their Laws*, Cartwright objected that we considered only tetrad equations and vanishing first order partial correlations. Why not higher order partial correlations, or other constraints besides tetrad equations? Good question. But, having said that, she argued that using other constraints

> … is bound to be wrong, since the very fact that makes an nth order partial correlation the relevant one – that is the fact that there are n other causes operating – also makes the both higher and lower-order ones irrelevant … what qualitative relations are relevant in the data depends on what causal structure is true; and each causal hypothesis must be judged against the data that are relevant for the structure they are, in fact, embedded in (1989, 79).

Once more, she was thinking about confirmation rather than search. If a variable X has exactly n direct causes among a set of other variables, hypotheses about vanishing correlations controlling for $k < n$ variables are extremely relevant: their falsity may tell us that X has more than k direct causes. Hypotheses about vanishing correlations controlling for $p > n$ variables will be irrelevant, but that observation only poses a problem about structuring a reliable search using tests of vanishing partial correlations, not an argument that no such search is possible.

In *Discovering Causal Structure*, I and my collaborators considered only vanishing first order partial correlations, for the good reason that we did not have a general algorithm for calculating the higher order partial correlations implied by arbitrary directed graphs representing linear models with independent errors, or even for acyclic directed graphs representing

"recursive" linear models with independent errors. Then, in 1989, we read Judea Pearl's book, *Probabilistic Reasoning in Intelligent Systems*, which had appeared the year before, and the lights came on.

In the early 1980s, a number of statisticians had formalized the relation between directed acyclic graphs and the vanishing partial correlations they imply in corresponding linear models with independent errors, and, more generally, between directed acyclic graphs and conditional independence. The crux of the connection was called the (local) Markov condition, and is a generalization of Reichenbach's notion of screening off. Formally, the Markov condition is simply a restriction on how directed graphs whose vertices are variables are to be paired with probability distributions over the space of possible joint assignments of values to the variables. A pair $\langle G, \Pr \rangle$, G a directed graph and Pr such a probability distribution, satisfies the Markov condition if and only if for each variable X represented by a vertex **X** in G, conditional on the parents of **X** (in G) X is independent of any set of variables, none of whose members are represented by vertices that are descendants of **X** in G. In linear models with normal distributions, conditional independence is vanishing partial correlation, and the Markov condition can also be reformulated for vanishing partial correlations even in non-normally distributed linear models with independent errors.

Pearl not only reviewed this work, he and his students did something of great importance: they used it to provide a fast algorithm to decide, for any directed acyclic graph and any conditional independence statement involving only variables represented by vertices in that graph, whether the Markov condition applied to the graph implies the conditional independence. The algorithm used a graphical property Pearl discovered and called d-separation, although to add confusion it is now sometimes called the (global) Markov property. It is straightforward to prove that the Markov condition is necessarily true of any system of functional dependencies among variables in which the exogenous variables (those of zero in degree in the graph) are independently distributed. So, with d-separation, we could now compute the vanishing partial correlations of any order implied by any directed acyclic graph, and hence by any linear recursive equational model with independent errors. (It later (in 1994) became clear that, in one respect, d-separation is a more generally applicable notion than the (local) Markov condition. I showed that the local Markov condition fails for non-recursive linear equational models with independent errors, and, in a much deeper effort, Peter Spirtes showed that d-separation necessarily holds of them.)

Pearl's book also explored further assumptions about the relations between graphical structure and conditional independence, in particular

the assumption, which we now call faithfulness, that all conditional independence relations in the probability distribution follow from the Markov condition applied to the graph with which the distribution is paired. For linear models, faithfulness is a generalization of Spearman's principle.

In his 1988 book, Pearl explicitly rejected the idea that the graphical structures he described might be used to describe any model-independent causal relations. From our work on linear latent variable models, I and my colleagues had a quite different view, and it proved fruitful. In 1990, Spirtes, Richard Schemes and I used d-separation and tests of conditional independence (or vanishing partial correlations) in an algorithm for constructing causal models from data, provided there are no unrecorded common causes of measured variables, and assuming the Markov condition is true of causal relations and the probabilities of variable values. We also suggested that related searches could be found for latent variable models. The search procedure was not really feasible, and faster algorithms were soon proposed by Pearl and his student, Thomas Verma, and by those of us at Carnegie Mellon. We were able to prove that, assuming faithfulness, our algorithms almost surely converged to the Markov equivalence class of the true structure, although the convergence, we now know, is not uniform. And we were able to prove that faithfulness holds almost surely of distributions on continuous variables satisfying the Markov condition, a result Chris Meek later extended to models whose variables have only finite sets of values. In 1991, in collaboration with Steve Feinberg, Chris Meek, and Elizabeth Slate, we introduced procedures for predicting the results of interventions given a graphical causal model without latent common causes. Spirtes subsequently generalized this work to predictions from only partially known models, with latent common causes. Based on this work, Pearl subsequently published rules for prediction that form a special case of Spirtes' results, and Pearl's work evolved to the elegant and useful formalism he presents in this issue. In 1992, Spirtes devised a more general search algorithm that works even when unmeasured common causes may be present, and proved a similar convergence theorem for it, and, in collaboration with Verma, a proof that the algorithm has a kind of completeness.

We assembled these and other results in a book, *Causation, Prediction and Search*, published in 1993, now much cited but never much read. It is already dated by subsequent developments about feedback models, Bayesian and other search procedures based on model scores, search procedures for feedback models, new prediction algorithms, results about search with sample selection bias, more general graphical representations, and more.[2] None of the work subsequent to 1993 seems to be known

to philosophers, but the book has been the subject of several criticisms, including four (counting her forthcoming book) by Cartwright.

We wrote in *Causation, Prediction and Search*:

> The basis for the Causal Markov Condition is, first, that it is necessarily true of populations of structurally alike pseudoindeterministic systems whose exogenous variables are distributed independently, and second, it is supported by almost all of our experience with systems that can be put through repetitive processes and whose fundamental propensities can be tested. Any persuasive case against the Condition would have to exhibit macroscopic systems for which it fails and give some powerful reason why we should think the macroscopic natural and social systems for which we wish causal explanations also fail to satisfy the condition. It seems to us that no such case has been made (1993, 64).

(We call a system one gets by marginalizing out some exogenous causes "pseudo-indeterministic".) The argument should have had the qualification that it applies to systems without feedback. The argument is not just by burden of proof: try to make a system without feedback that violates the Markov condition When you take account of all causes, including your own actions. You won't succeed with hammer and nails, or light bulbs, batteries and wires, or household chemicals, or your computer. It is easy to produce apparent but inauthentic counterexamples: collapse the values of variables into a reduced set, and ignore that you have collapsed them; marginalize out common causes and ignore that you have done so; mix together systems with different propensities and average the probabilities over them, while ignoring that you have mixed and that the systems are of different kinds; draw a sample by a method that determines membership in the sample by the values units have for two causally unconnected variables, and ignore how the sample was obtained. There are, of course, physical phenomena, such as hysteresis, in which proximate state descriptions do not screen off prior state descriptions. I think, that these are the very circumstances in which we ordinarily infer that the state descriptions are incomplete. Ignorance can easily produce apparent counterexamples to the Markov condition as a causal principle, and in inquiry we are often in ignorance, but the ignorance should not be willful. Genuine counterexamples may exist, but they are not ready to hand.

Cartwright has risen to the challenge with an example she has repeated in several places. I quote from the manuscript of her forthcoming book.

> Consider a simple example. Two factories compete to produce a certain chemical that is consumed immediately in a nearby sewage plant. The city is doing a study to decide which to sue. Some days chemicals are bought from Clean/Green; others from Cheap-but-Dirty. Cheap-but-Dirty employs a genuinely probabilistic process to produce the chemical. The probability of getting the desired chemical on any day the factory operates is eighty percent. So in about one-fifth of the cases where the chemical is bought from Cheap-but-Dirty, the sewage does not get treated. But the method is so cheap the city is prepared to put up

with that. Still they do not want to buy from Cheap-but-Dirty because they object to the pollutants that are emitted as a by-product whenever the chemical is produced.

That is what really is going on, but Cheap-but-Dirty will not admit to it. They suggest that it must be the use of the chemical in the sewage plant itself that produces the pollution. Their argument relies on the screening-off condition. If the factory *were* a common parent, C, producing both the chemical X and the pollutant Y, then (assuming all other causes of X and Y have already been taken into account) conditioning on which factory was employed should make the chemical probabilistically independent from the pollutant, they say ... Cheap-but-Dirty is indeed a cause of the chemical X, but they cannot be a cause of the pollutant Y as well they maintain since

$$\text{Prob}(X, Y \mid C) = 0.8 \neq 0.8\ X\ 0.8 = \text{Prob}(X \mid C) \bullet \text{Prob}(Y \mid C)$$

(Cartwright, forthcoming)

The story seems incoherent on its face: if, as Cheap-but-Dirty claim, it is the use of the chemical in the sewage plant itself that produces pollution, then by the Markov condition assumed in Cheap-but-Dirty's explanation (the tool of evildoers everywhere) pollution should be independent of which factory produced the chemical, but that independence is implicitly denied at the beginning of the story.

That aside, consider where this factory is: *nowhere.* Cartwright claims to refute a hypothesis about nature, the Causal Markov condition, by *imagining* a counterexample. We can refute the special theory of relativity in the same way – imagine a positive rest mass accelerated from less than the velocity of light to more than the velocity of light. Our scientific ancestors could have refuted the impossibility of a perpetual motion machine in the same way; no need to wait for the discovery of Brownian motion. They say rabbits don't have opposable thumbs, well just imagine otherwise. Perhaps we can catch them by those thumbs.

Cartwright has another argument against the Causal Markov condition. Given any system of causal relations, all sorts of probability distributions on the variables, including distributions violating the Markov condition, are mathematically possible, hence conceivable.

Nothing in the concept of causality, nor of probabilistic causality, constrains how Nature must proceed ... Lesson: where causes act probabilisticaily, screening off is not valid (Cartwright, forthcoming)

Parallel arguments: Nothing in the concepts of light speed, rest mass, velocity and acceleration constrains how Nature must proceed. Lesson: the special theory of relativity is not valid. Nothing in the concepts of rabbit or of opposable thumbs constrains how Nature must proceed. Lesson: it is not true that rabbits do not have opposable thumbs.

That aside, I don't know how Cartwright (or anyone else) knows what is or not *in* the concept of causality. That seems more a psychological than a philosophical matter. For all I know (and, I'll wager, for all anyone knows), infants may be born with a tacit understanding of causality that requires the Markov condition, or they may learn about what causes what so in train with instances of the Markov condition that the two form one concept in the sense that in simple cases the inference from sequences of causes to conditional independence is automatic. What is automatic is not necessarily articulate.

Rabbit thumbs are close to the right example. I think Cartwright's difficulties with the Causal Markov condition come, on the one side, from the fact that it is not an explicit theoretical principle, like the limiting velocity of light, that is part of a general, well tested, exact theory about the structure of the universe. On the other side, neither does the principle have some a priori justification that would establish that rationality requires acting as if it were true. Save that it is far more general, the Markov condition is more like the proposition that rabbits don't have opposable thumbs. Conceivably, by chance, now and then a rabbit does have an opposable thumb, but its an anomaly. Conceivably, we could do some ugly genetic work to create a rabbit with opposable thumbs. But trying to catch rabbits by the thumb would be a very bad strategy because (this time, because), they almost never have thumbs.

The remainder of Cartwright's forthcoming objections cannot be usefully summarized or quoted. They are a variety of arguments, based in part on her denial of the Markov condition, that the methods in *Causation, Prediction and Search*, and other automated search methods over graphical models, cannot save in exceptional circumstances succeed in finding causal relations. She kindly says that our theory and methods are "very beautiful", while denying they are of much use. She recommends instead "hypothetico-deductive method" adapted to "individual circumstances". Either the Markov condition is false, she argues, or if true, other features of nature (for example the mixture of causal structures in natural populations) make the methods uninformative. We have considered her arguments for the first horn of this supposed dilemma, what about the second?

I believe that Cartwright has never used the program whose theory *Causation, Prediction and Search* documented, has never analyzed any real data with it, and has never read the manual that goes with it. (Given her views she has little motive.) Else she would not contrast our methods with those adapted to "individual circumstances". Applying the program to real data requires a lot of adaptation to particular circumstances: variables must often be transformed to better approximate normal distributions, decisions

made about modeling with discrete or continuous variables, data must be differenced to remove auto-correlation, and on and on. The program allows the user to specify a range of assumptions adapted to the "individual circumstances": latent variables can be allowed or forbidden, and particular causal connections can be forbidden or required.

I will give five examples of positive causal information produced by the procedures, cases where, either by independent interventions or by well established independent knowledge not used in the data analysis, predictions of the procedure were established.

Case 1. Using data from the U.S. News and World Report surveys of American colleges and universities, Druzdzel and Glymour predicted in 1993 that increasing the average SAT scores of the Freshman class at a college or university would reduce the dropout rate. The recommendation was put into practice at Carnegie Mellon by altering the formula for scholarship support. Average SAT scores of entering Freshmen improved in each subsequent year, and dropout rates decreased accordingly.

Case 2. Using a small sample (n = 45) of observational data collected in the 1970s, and published in a recent textbook on regression, I and my collaborators predicted (retrodicted really) that of 14 measured variables, only one, pH, directly influenced the biomass of Spartina grass growing in the Cape Fear estuary. That is, if pH were held constant over the range of values exhibited in the data, variation in other variables (throughout the range exhibited in the data) would not influence biomass. The prediction was contrary to regression analyses, which, as reported in the textbook, gave a variety of different results depending on the regression technique used. Our prediction was also contrary to the prediction of the biologist who collected the data, and who believed that sodium ion concentration – salinity – also influenced biomass directly.

Subsequent to the prediction, we obtained a copy of the doctoral thesis which contained the original data. The thesis reported a subsequent randomized block greenhouse experiment with plugs of Spartina grass from the same estuary. The experimental treatments varied pH, sodium ion concentration and mechanical aeration, and biomass, measured as in the observational study, was the outcome. Ph directly influenced biomass, but the other two variables had no effect when pH was held constant.

Case 3. Data from a Swedish satellite with a mass spectrometer, intended to measure concentrations of a variety of light and heavy ions in space, violated well established concentration ratios among the ions, presumably due to a known miscalibration of the instrument before launch and

miscalibrations resulting from the space environment. Using our program, physicists at the Swedish Institute for Space Physics concluded that the instrument reliably measured total concentrations of heavy ions and total concentrations of light ions, but not concentrations of particular species. After recalibration of the data interpretation software, the differences from theory were reduced by half

Case 4. Given data for seven psychometric tests, and a total score (AFQT score) obtained by averaging some or all of the test scores (we were actually told by the person who provided us the data that all seven scores were used in calculating AFQT) together with other test scores not in the data, a regression analysis predicted that two of the 7 psychometric scores had not been used in forming the AFQT score for each person in the data set. Our program predicted that four of the psychometric test scores had not been used in calculating AFQT, and, of course, which four. Our prediction was correct.

Case 5. Recently, I have been working with the National Aeronautics and Space Administration Ames Research Center on software to identify the composition of rocks and soil from reflectance spectra. The aim is to enable future planetary rovers to interpret spectral data themselves. After a year of investigation using simulated data from libraries of spectra of pure minerals, we obtained the spectrum of a rock of unknown composition from the Mojave desert. A variety of regression procedures produced equivocal results about the composition of the rock. Using our programs, and assuming the minerals composing the rock were from among 135 minerals in a library of spectra (an assumption that turned out to be correct), we predicted the rock was composed of calcite and dolomite. Chemical and microscopic examination of the rock at Washington University subsequently reported it was dolomite with calcite veins. Whether there are other minor minerals in the rock is not yet known. In a more extensive test, a simplied (for *the individual circumstance*) version of the PC algorithm described in *Causation, Prediction and Search,* was tested against an expert human geological spectroscopist. The test consisted of deciding whether certain mineral classes were present in 191 rock samples of known composition, given only the reflectance spectrum of each rock. There were 17 possible mineral classes to be identified, of which 7 did not occur or occurred in 10 or fewer of the rock samples The human expert had unlimited time (he took about 12 hours), could and did consult reference works, and had prior knowledge of the number of minerals of each class present in the 191 samples. For all but one class of minerals (nesosillicates) the

automated procedure performed comparably to the human expert, in some respects slightly worse, in other respects slightly better. This is a problem in which the data are produced by a mixture of different processes, and almost certainly a non-linear mixture. In subsequent tests in which the task was specifically to identify carbonaceous rocks and soils, the algorithm performed substantially better than a human expert.

In many cases, especially with social data sets, the programs give negative information. Where the programs cannot make reasonable inferences about causal explanations, they tell the user as much: either a program produces a spaghetti of connections and tell the user it cannot determine whether the connections are produced by direct causal influences, or by latent common causes (or mixtures of structures), or both, or the results vary a great deal with small variations in the significance levels used in tests of conditional independence, or the program simply never comes back with a result. In cases such as these – the data used in *The Bell Curve* is a typical example – the methods we use say that nothing can be inferred from the data about causal relations. In those cases, the program tells us that an unlimited number of models can explain the data, consistent with the assumptions given to the program by the user. Cartwright seems to think this sort of negative information is useless, presumably because she thinks in such situations social (and other) scientists can do better, they can somehow find the truth from the data rather than impose their prejudices upon it. How? Improved search algorithms or flexibility about the statistical tests used, or the use of Bayesian scores versus constraints founded on hypothesis tests, all of these may yield improved methodology. But they are clearly in the spirit of the work in *Causation, Prediction and Search*, and indeed are part of the later research it helped to provoke. One might sometimes do better by guessing a true structure, unfaithful (or close enough given the sample size) to the probability distribution estimated, but guessing is not a method. One might do better by having prior knowledge about the systems under study (as our geological expert did), but if that knowledge is articulate it can be put into the machine, and that process was discussed and illustrated in *Causation, Prediction and Search*.

3. CONCLUSION

What is Cartwright's method? Cartwright's first and second books proposed methods for discovering causal relations that I cannot clearly describe. *Nature's Capacities and Their Measurement* contrasts her methods with mine (and my collaborators, in 1987); ours are said to be

"hypothetico-deductive", which is in some sense true, but chiefly indicates the poverty of philosophical vocabulary in talking about search. I do not have enough of her forthcoming book to know what methods, if any, it advocates, but the chapter I have read suggests methods of inquiry must now be "hypothetico deductive". But hypothetico-deductive method is not a *method* of inquiry; it is at most a cog in a method, and as conventionally used in social statistics, where a few guesses are tested and all other possible guesses ignored, not even that.

The majority of *Nature's Capacities and Their Laws* is really devoted to developing a metaphysical conception of probabilistic causation that I think is perceptive and has proved fruitful. I have not described Cartwright's positive metaphysic because, no matter the offence it may give, this essay is defensive. But I wish to praise her metaphysic. By my lights, philosophy of science should be largely judged by its contribution to scientific progress, and by that measure *Nature's Capacities and Their Laws* stands out. Patricia Cheng recently proposed a psychological model of human judgement, justified by an impressive array of experimental results. The picture of causation Cheng employs is Cartwright's, and although Cartwright is not cited, Cheng tells me she had read *Nature's Capacities and Their Laws*, and may have forgotten an influence. Cheng's empirical work is accompanied by some ingenious algebra that shows, much as Cartwright might have wished, that what Cheng calls "causal powers" and Cartwright "capacities" can sometimes be identified from probabilities given certain sorts of prior knowledge, even when an effect may also have unobserved causes. In view of Cartwright's claim that this part of her book was intended to give the philosophical reasons for a metaphysical mistake I am supposed to have made, I take a certain pleasure in the fact that Cheng's models turn out to be isomorphic to a particular parameterization of directed graphical models, the very structures for which we search, and that applying simple formal techniques for graphical models, not least the Markov condition yields new predictions from her theory.[3]

NOTES

[1] There is a complex sense in which the models are not equivalent. In the model on the right ρ cv,en.po can be made to vanish by adjusting the dependencies among cv, T, and po so that cv and po are uncorrelated. That cannot by done in the model on the left.

[2] Links to much of this later work can be found on the web pages of David Heckerman, of Chris Meek, of Judea Pearl, of Thomas Richardson and of Peter Spirtes.

[3] See Glymour 'Psychological and Normative Theories of Causal Power and the Probabilities of Causes', *Proceedings of the 1998 Conference on Uncertainty in Artificial*

Intelligence, and Cheng, P. W.: 1997, 'From Covariation to Causation: A Causal Power Theory', *Psychological Review* **104**, 367–405.

REFERENCES

Cartwright, N.: 1979, 'Causal Laws and Effective Strategies', *Nous* **13**, Reprinted in Cartwright (1983).
Cartwright, N.: 1983, *How the Laws of Physics Lie*, Oxford University Press, Oxford.
Cartwright, N.: 1989, *Nature's Capacities and Their Measurement*, Oxford University Press, Oxford.
Glymour, C., R. Schemes, P. Spirtes, and K. Kelly: 1987, *Discovering Causal Structure*, Academic Press, San Diego, CA.
Glymour, C., R. Schemes and P. Spirtes: 1993, *Causation, Prediction and Search*, Springer-Verlag, Vienna and New York.
Pearl, J.: 1989, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, AP Professional, Chestnut Hill, MA.

Department of Philosophy
Carnegie Mellon University
Pittsburgh, PA 15213
USA
E-mail: cg0g@andrew.cmu.edu

Department of Philosophy
University of California, San Diego
9500 Gilman Drive
La Jolla, CA 92093-0119
USA