**Reply to Humphreys and Freedman's Review of Causation, Prediction, and Search**

Peter Spirtes; Clark Glymour; Richard Scheines

*The British Journal for the Philosophy of Science*, Vol. 48, No. 4 (Dec., 1997), 555-568.

Stable URL:
http://links.jstor.org/sici?sici=0007-0882%28199712%2948%3A4%3C555%3ARTHAFR%3E2.0.CO%3B2-A

*The British Journal for the Philosophy of Science* is currently published by Oxford University Press.

*Brit. J. Phil. Sci.* **48** (1997), 555–568

DISCUSSION

# Reply to Humphreys and Freedman's Review of *Causation, Prediction, and Search*

Peter Spirtes, Clark Glymour, and Richard Scheines

In an essay in this journal entitled 'The Grand Leap' (Humphreys and Freedman [1996]) Paul Humphreys and David Freedman have offered a highly critical review of our book *Causation, Prediction, and Search* (Spirtes, Glymour, and Scheines [1993];[1] henceforth CPS). By omission and commission, their essay repeatedly and systematically misdescribes what we wrote, so much so that it is impossible for a reader of their article to glean even the most basic understanding of what we claimed. Their review is riddled with false claims about what the procedures we described can and cannot do, and fundamental technical errors. Many of the objections Humphreys and Freedman lodge against us would equally be objections to methods of causal inference (e.g. randomized clinical trials) that are universally accepted; still other of their 'criticisms' are simply repetitions, without attribution, of cautions we made in the book they purport to review. In still other cases Humphreys and Freedman contrast our work with efforts they think better when, in fact, the work they praise derived directly from *Causation, Prediction, and Search*, and that legacy is explicitly acknowledged by the authors.

What we did in *Causation, Prediction, and Search* was straightforward. We used an existing formalism—parametrized directed acyclic graphs, sometimes known as Bayes networks—to represent both the causal claims and the probability constraints of otherwise diverse classes of statistical models that are used in causal explanations for continuous and categorical data. Under a variety of explicit, formal assumptions relating causal hypotheses (in the form of directed graphs) to constraints on associated probability distributions, we characterized statistical indistinguishability; that is, we showed how to decide whether two or more alternative causal explanations are indistinguishable from probabilities or from constraints on probabilities.

[1] CPS is currently out of print, but is available on the World Wide Web at the address http://hss.cmu.edu/html/department/philosophy/TETRAD.BOOK/book.html.

(Some of these indistinguishability results were already in the statistical and computer science literature; see e.g. Verma and Pearl [1990].) Then, using two explicit assumptions relating systems of causal hypotheses with probabilities, we showed that there are algorithms that, provably, in the large sample limit find causal features common to the equivalence class of the causal structure that generated the data. Then we tested the algorithms on (i) randomly generated data from randomly generated structures (randomly parametrized random graphs); (ii) randomly generated data from structures elicited from experts; (iii) data from empirical cases where causal features were known independently, and (iv) empirical data where published causal explanations were in the literature but the true explanation was not known. Then we gave (again, provably correct under explicit assumptions) procedures for calculating, from partial causal information and marginal probability distributions, the effects of interventions. Finally, we began the investigation of properties that follow from both strengthened and weakened assumptions connecting causation and probability, including the case of feedback systems. The reader of Humphreys and Freedman's review would discover almost none of this. We turn to their criticisms and misrepresentations.

1. In CPS, we wrote: 'We advocate no definition of causation, but in this chapter attempt to make our usage systematic, and to make explicit our assumptions connecting causal structure with probability, counterfactuals, and manipulations' (CPS, p. 41). We had two reasons for offering no definition. First, we know of no satisfactory reductive definition of causation. Second, none of the inferences that we make depended upon having a reductive definition of causation—instead they depend only upon certain relationships between causal structures and probability distributions which we introduced axiomatically.

Humphreys and Freedman say:

> SGS do not give a reductive definition of 'A causes B' in non-causal terms. And their axiomatics require that you already understand what causes are. Indeed the Causal Markov condition and the faithfulness assumption boil down to this: direct causes can be represented by arrows when the data are faithful to the true causal graph that generates the data. In short, causation is defined in terms of causation. That is why the mathematics in SGS will be of little interest to philosophers seeking to clarify the meaning of causation (p. 116).

and they added in a footnote:

> SGS justify their lack of an explicit definition by noting that probability theory has made progress despite notorious difficulties of interpretation—perhaps the first innocence-by-association argument in causal modelling. On the other hand lack of clarity in the foundations of statistics may be one source of difficulty in applying the techniques (*ibid.*).

Humphreys and Freedman do not deny—or note—that there are scientific fields in which progress is made despite disagreement over definitions of key terms (for example: 'point' in geometry, 'force' in Newtonian physics, 'set' in set theory, 'probability' in probability theory, etc.). Is there some particular special fact about causal inference which, unlike other fields, would prevent all progress until a satisfactory, generally agreed-upon reductive definition of cause is given? They offer no such argument. Are they seriously proposing that all work on statistics stop until we all agree on its foundations? Should we abandon most of statistical inference, epidemiology, and randomized clinical trials until we all agree on a reductive definition of 'causality'?

In a footnote, Humphreys and Freedman contrast what they consider to be our flawed approach with other approaches that do give a 'formal treatment of causation, in the sense of effects of hypothetical interventions' (p. 115, footnote 7). One of the sources that they contrast with our approach is Pearl [1995]. Humphreys and Freedman do not tell the reader that there is one source that Pearl cites as the origin of what Humphreys and Freedman call Pearl's 'formal treatment' of hypothetical interventions—namely, *Causation, Prediction, and Search*.[2]

2. We discuss several assumptions relating causation and probability and investigate the consequences of two of them in detail, those we call the Causal Markov and Faithfulness Conditions. We use directed acyclic graphs (DAGs) to represent causal relationships between variables. For example, suppose there are two variables A and B such that A does not cause B, B does not cause A, and there is no third variable which causes both of them. In that case, the DAG that represents the causal structure of A and B is the empty graph (i.e. no directed edges) shown in (i) of Figure 1. Similarly, A causes B is represented by (ii), and B causes A is represented by (iii).

The two assumptions serve to associate with each causal DAG a set of conditional independence and dependence relations.[3] For example, in (i) of Figure 1 the Causal Markov and Causal Faithfulness Assumptions entail that A

| A B | A → B | A ← B |
|---|---|---|
| A not cause B<br>B not cause A | A cause B<br>B not cause A | A not cause B<br>B cause A |
| (i) | (ii) | (iii) |

Fig. 1. Three causal graphs

[2] According to Pearl [1995], 'Much of this investigation was inspired by Spirtes et al. (1993), in which a graphical account of manipulations was first proposed.'

[3] Here, the phrase 'conditional independence relations' is intended to include independence relations as a special case, and similarly 'conditional dependence relations' includes dependence relations as a special case.

is independent of B; in (ii) and (iii) of Fig they entail that A and B are dependent. Of course, in more complex examples, the set of conditional independence and dependence relations they entail is considerably more complex.

In CPS we discuss the arguments for the Causal Markov and Faithfulness Conditions at length, explain why they are plausible assumptions in many (but not all) domains, and investigate a variety of ways in which each of them can fail, or appear to fail—including feedback systems, deterministic relations, and populations with units having distinct causal structures. Here we will simply note that they are very widely (if implicitly) assumed in statistical and experimental reasoning, for example in the design and interpretation of randomized clinical trials.

According to our critics,

> the Causal Markov condition and the faithfulness assumption boil down to this: direct causes can be represented by arrows when the data are faithful to the true causal graph that generates the data. In short, causation is defined in terms of causation (p. 116).

We take this to mean that they accuse us of taking satisfaction of the Causal Markov and Faithfulness Conditions to be part of the meaning of 'cause'. Note that they do not offer *any* text from CPS to support this charge—the only text that they quote *flatly contradicts* it ('We advocate no definition of causation,' CPS, p. 41). Further text that they do not cite also flatly contradicts this claim ('The Markov condition is not given by God; it can fail for various reasons we will discuss in the course of this book. The reliability of inferences based upon the Condition is only guaranteed if substantive assumptions obtain.' CPS, p. 9). This tactic—to attribute an absurd view to us without citing any text whatsoever to justify the attribution—is one that they have used repeatedly.[4] In this case, they also left out passages that flatly contradict the view that they attribute to us.

Remarkably, in all of their complaining about the Causal Markov and Faithfulness Conditions, Humphreys and Freedman offer not one single argument against the *truth* of these assumptions. They do complain that assumptions have been made ('Thus causation is not proved into the picture, it is assumed in,' p. 115). In most statistical or experimental design papers, assumptions relating causal structures to probability distributions are not

[4] For example, they criticize us for adopting something they call the Automation Principle (p. 121). We have never advocated such a principle, it plays no role in any argument that we have ever made, and they offer not one single citation justifying their attribution of this principle to us. It also ignores the fact that all of the algorithms allow users to input any background knowledge about time order, or whether or not one variable causes another variable. They also say (p. 114) that we claim that our search methods are superior to path models and hierarchical linear models. We made no such claim, and indeed the claim is nonsensical, since it is meaningless to compare a search algorithm with a probabilistic model.

stated explicitly, nor are their limitations explored, as we did. But it is obvious that anyone who draws causal inferences from statistical data has to make *some* assumptions relating causal structures to probability distributions.

3. A number of different search algorithms are analysed in CPS, algorithms which make different assumptions and have different outputs. It is often difficult to tell which algorithm Humphreys and Freedman are discussing, since they never say. (The confusion is compounded by the fact that they repeatedly criticize 'TETRAD' (e.g. pp. 116–17), which is a program which we wrote in 1987 and has nothing to do with the algorithms they are discussing, which are implemented in a program called 'TETRAD II'.) Some of the statements that they make about 'the algorithms' are true of only some of the algorithms; others that they make (e.g. that we use t-tests, p. 117) are not true of any of the algorithms, as CPS makes clear. Two of the principal algorithms discussed in our book are the PC procedure, which assumes there are no unmeasured common causes of recorded variables, and the FCI algorithm which dispenses with this limitation. Anyone reading the review would be led to believe that the output of both the PC and FCI algorithms are single graphs. They say: 'The object of the SGS algorithms is to reconstruct the graph from these statistical relations' (p. 115). They also say: 'TETRAD (*sic*) also orients the edges that remain' (p. 117). This is all false. If we were to output a single graph, then given a single pair of dependent variables X and Y, we would have to arbitrarily choose the single model 'X causes Y' or arbitrarily choose the single model 'Y causes X', when the data say nothing about which of these is the correct model. For that reason, the outputs of the PC and FCI algorithms are not graphs but objects (patterns and POIPGs respectively) that represent sets of graphs, a fact emphasized over and over in CPS and illustrated with numerous diagrams. Thus given a single pair of dependent variables X and Y, the PC algorithm outputs a graphical object that represents the disjunction X causes Y or Y causes X by placing an *undirected* edge between X and Y. In some cases the set of graphs represented by the PC output is a singleton, but the FCI output (without background knowledge) is *never* a singleton. Whenever there is more than one causal model in the set of causal models represented by a pattern of a POIPG, not all of the edges are oriented. The output of the FCI algorithm (without background knowledge) *always* contains some edges that are not oriented.

4. Humphreys and Freedman complain that CPS only describes procedures for inference with independently, identically distributed (i.i.d.) samples. This is false. Statistical tests for conditional independence in such cases are not developed, but what can be inferred from conditional independence in such cases is extensively discussed in the book. Non i.i.d. samples can arise, for example, when the sample is drawn from a mixture of two or more distinct populations with distinct probability distributions. Chapter 6 of the book

describes the causal inferences possible from conditional independence facts in mixed samples. Non i.i.d. samples also arise in experimental designs (as in clinical trials) in which the treatment of later subjects depends on the outcomes from earlier subjects. Such cases are discussed in Chapter 9. Since writing CPS, we have shown how in general, under a slight re-interpretation of the output of the FCI algorithm, causal connections between the measured variables and the property of inclusion in the sample do not affect the correctness of the algorithm's output (although they may certainly make the output very uninformative in the sense that the output is a disjunction of many causal models that share no interesting features in common. See Spirtes *et al.* [1995]).

5. Implying that the algorithms we describe will not work if variables are measured with error, Humphreys and Freedman claim that 'the Markov Condition must hold for the original variables to which the algorithms will be applied; it is not enough if it holds for recoded variables' (p. 115). This is true. However, it does *not* imply that the correctness of the FCI algorithm depends upon how the variables are coded. While the *informativeness* of the output of the FCI algorithm depends upon how variables are coded, the correctness of the algorithm does not.[5]

6. Humphreys and Freedman complain that the TETRAD II programs only apply to two families of distributions, the normal and the multinomial. The normal family is almost coextensive with linear models, and so it is very odd that Humphreys, who has claimed that *all* causal relations are linear (Humphreys [1989]), should make such an objection. In fact—a fact one would not discover in their 'review'—the TETRAD algorithms are modular. An inference procedure calls an oracle for information about conditional independence, and uses that information to search for causal explanations. The oracle can be any reliable source of information about conditional independence relations among measured variables. In the fully automated parts of the program, the oracle is implemented by statistical tests based either on the normal or multinomial distributions, but the TETRAD II program contains procedures for users to introduce conditional independence facts from whatever source, which are then used by the inference procedures. Expanding the automated oracle to include certain families of distributions, for example the conditional Gaussian, is a programming exercise. In other cases there is statistical work to be done in finding useful tests of conditional independence.

7. In their review, and other articles, Humphreys and Freedman have

[5] If, for example, blood pressure is only recorded as 'high', 'normal', or 'low', we can represent blood pressure (as measured by two numbers representing diastolic and systolic blood pressure) as a latent cause of measured blood presure. This introduces some deterministic relations between variables, which slightly complicates the proof of correctness, but the machinery developed in Chapter 3 of CPS can be used to handle this case.

repeatedly insinuated that we have deliberately misled readers about what the algorithms can and cannot do. They claim that we are 'exaggerating' (p. 113), that the book is 'show business' (p. 123), that we 'seem to offer empirical proof' (p. 118) but the 'proof is illusory' (p. 118), that no matter what the output of the algorithm we 'count a win either way' (p. 119), and that some of the assumptions are 'not emphasized' (p. 116) even though they appear in the computer output and documentation and the book itself. The reader should note that they do not charge that *any* of our claims about the evidence for the reliability of the algorithms were false. Moreover, as we will document below, in CPS we went out of our way to indicate conditions under which the algorithms were not reliable or produced bad output. Humpreys and Freedman systematically fail to mention any of these statements.

Humphreys and Freedman claim that much of our evidence about the reliability of the algorithms is 'illusory' because the theorems about correctness in the limit are not informative about the performance of the algorithms on realistic sample size, some of the evidence comes from simulated data, and because we discuss some hypothetical cases. What they do not tell the reader is that in every case simulated data are clearly labelled as simulated, and that the hypothetical models are clearly labelled as hypothetical.

Humphreys and Freedman (p. 117) say about the theorems of correctness we prove:

> However, it is exact independence that is relevant, and exact independence cannot be determined from any finite amount of evidence. Consequently, the mathematical demonstrations in SGS (e.g. Theorem 5.1 on p. 405) do not cope with basic statistical ideas. Even if all the assumptions hold, the t-test makes mistakes. Therefore, the SGS algorithms can be shown to work only when exact conditional independencies and dependencies are given.[6]

Perhaps they would prefer, as with almost all statistical search procedures (e.g. factor analysis, stepwise regression, modification indices), that no proof of correctness be available. We proved that, given our assumption, with an oracle that can correctly answer questions about conditional independencies and dependencies in a population, the outputs of our algorithms are correct. One way to realize such an oracle in the large sample limit is to perform statistical tests of conditional independence. Humphreys and Freedman point out that this kind of theorem does not guarantee success on finite samples. We agree; that is why we said on the page following Theorem 5.1 (CPS, p. 115): 'We need to consider whether an algorithm remains reasonably reliable when the data are imperfect.' Of course, given the problem of sampling error, *no* algorithm whose output is a function of the sample could guarantee success.

[6] HF are wrong when they claim the algorithms use t-tests. None of the algorithms that we implemented uses t-tests; the tests we do use are clearly described in Chapter 5.

Once again, the same kind of charge that Humphreys and Freedman bring against TETRAD II could also be made about standard experimental design. Tests of difference of means, for example, commonly used in analysis of experimental outcomes, are tests of consequences of independence, and can only be guaranteed to give correct answers in the large sample limit.

Because we realized the need to test the algorithms on finite sample sizes, we performed a series of simulation studies. We randomly generated directed acyclic graphs, and random parametrizations of the graphs, and randomly generated samples of various sizes from the probability distributions characterized by the parametrized graphs. We then gave the samples to the algorithms, and asked the program to reconstruct the graphs. In these simulations we found that in large samples (more than 2000), where each variable had relatively few parents (two or three), the algorithms were highly reliable at correctly finding adjacencies; they were less reliable at finding orientations. When variables had more parents (4 or 5) both the adjacencies and the orientations output by the algorithm were much less reliable.

Simulation studies are an important standard tool for putting an upper limit to the reliability of an algorithm and seeing how sampling error affects the output, widely used throughout statistics. It would have been irresponsible for us not to do them. Of course, it is important to clearly label the simulation studies as simulation studies, and to point out the limitations of simulations studies. We took both of these steps. In addition to looking at simulated data from randomly generated graphs, we looked at simulated data from a structure called the 'Alarm network', constructed by medical experts. This network has been used to evaluate the performance of many different search algorithms, and so provides a convenient benchmark for comparison (see e.g. Cooper and Herskovits [1992], Chickering *et al.* [1994]). Once again, we clearly stated that the data were simulated. Once again, it is difficult to see what Humphreys and Freedman have to complain about. Are they suggesting that no one should do simulation studies any more, no matter how clearly they are labelled?

In CPS, we also discussed a number of hypothetical examples, all of which were clearly labelled hypothetical. Do Humphreys and Freedman think that any discussion of hypothetical examples, no matter how clearly labelled, is illegitimate? In the past, in criticizing our work, Freedman himself has discussed hypothetical examples with data he made up (Freedman [1994]); but for some unexplained reason this is a legitimate tactic when Freedman employs it. The most famous work on experimental design, Fisher's *The Design of Experiments* [1960], begins with an imaginary example.

In discussing the empirical examples in CPS, Humphreys and Freedman say:

> What are the scoring rules? Apparently, SGS count a win if their algorithms more or less reproduce the original findings (rule no. 1), but they

> also count a win if their algorithms yield different findings (rule no. 2). This sort of empirical test is not particularly harsh (p. 119).

(After accusing us of counting a win no matter what, in a footnote to this passage which belies the passage itself, Humphreys and Freedman admit that as a matter of fact we do acknowledge shortcomings of the procedures.) The only basis for this remark is that we compare the program output for data from several social scientific studies—in which no one knows the true causal process—and in some cases the program agrees with the published models and in other cases it proposes alternative explanations. But in the empirical examples we evaluated the output of the algorithms in a number of different ways that are not guaranteed to produce success no matter what the output, including the following:

i. In two cases (Spartina biomass and AFQT) we compared the output of the program to the part of the causal structure known independently.
ii. We recommended that if whether or not two variables were adjacent in the output of the program depended heavily upon significance level, then conclusions about whether those edges were there or not should be considered suspect. In the case of Spartina biomass, we explicitly pointed out that 'there is one robust conclusion', namely that Ph was the controlling cause of the Spartina grass biomass (which was partially confirmed by experiment.)
iii. A relevant feature of the remarks made in the book about the Blau and Duncan, and about the Duncan, Featherman, and Duncan studies of occupational mobility, is that in these cases the TETRAD II program, run under the assumption that there are no latent variables, produces a complete, or nearly complete graph. As we point out, such output means that the program cannot really determine whether or not the associations are due to unrecorded common causes or mixtures, because a structure with a system of unrecorded common causes or mixtures of nonlinear structures will produce such a graph.
iv. In the case of the Weisberg rat liver data, we estimated (using simulation techniques) the probability of type II error of the algorithm against a specific alternative causal structure. That is, when the search procedure output a model M on the actual data, we investigated how often the search procedure would output M even on data generated by a specific alternative structure $M'$.

In our discussion of the simulation study results the scoring rules we use are described explicitly, and one of the conclusions drawn is that the orientations output by the program are very unreliable for degree greater than 3 even at reasonably large (e.g. 2000) sample sizes. We also extensively discuss when

the assumptions made by the algorithms may fail, and what kinds of mistakes errors in judgements of conditional independence can make.

Now to the Rindfuss *et al.* [1980] study of education and fertility, about which Humphreys and Freedman insinuate we were deceiving our readers. They claim that the graph on the left in Figure 2 is what we reproduced in CPS, and the graph on the right is the complete output. (We explicitly noted in CPS that we were reproducing only part of the output.) However, the graph on the right is not the output produced by the commercial version of the program. (We are not sure why the discrepancy exists; perhaps they had a beta test version of the program.) The output of the commercial version of the program differs from the output that Humphreys and Freedman claim in two important respects we explain below—it has additional adjacencies, and a number of double-headed arrows.

The first point to note is that all of the claims that we made about the Rindfuss case are true, and not disputed by Humphreys and Freedman. The *only* claim that we made about the Rindfuss example was that using the same assumptions as Rindfuss *et al.* (no latent variables, linearity, and a time ordering of certain subsets of variables) the model output by the algorithm was similar to the structure that they had argued for from background knowledge. We *explicitly* remarked that in general for the empirical cases from the social sciences we did not know the true model (CPS, p. 133).

Rindfuss *et al.* were interested in estimating what they thought was a reciprocal causal relation between mother's age at birth of first child (AGE), and her education (ED). They did a two-stage, least-squares regression, which in this case required that they find some regressor that was not a cause of AGE, and some other regressor uncorrelated with the first regressor that was not a cause of ED. They argued on substantive grounds that father's occupation

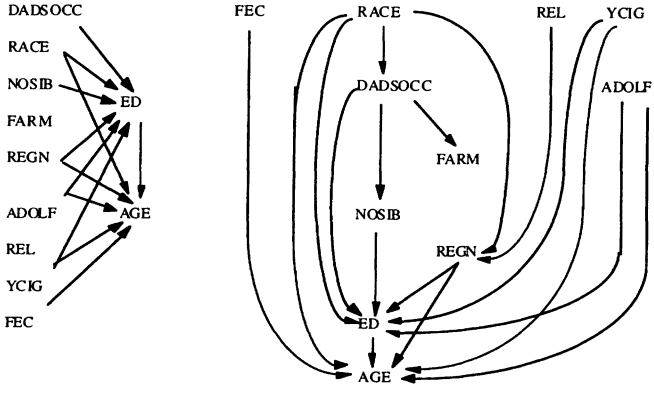


Fig. 2. Education and fertility

(DADSOCC) did not cause AGE, and whether the subject had a miscarriage (FECUND) did not cause ED. (In addition, FECUND and DADSOCC are uncorrelated.) After estimating what they thought was a reciprocal effect of ED and AGE on each other, they found that ED affected AGE, but not the other way around. Given the assumption of linearity, and a partial time ordering of the variables (that is, that AGE and ED are not causes of the other variables), TETRAD II reproduced the substantive claims that DADSOCC has no direct influence on AGE, FECUND has no direct effect of ED, and ED causes AGE but not the other way around.

The relations among the regressors are completely irrelevant to the claim that we made. Because Rindfuss *et al.* performed a two-stage, least-squares regression analysis of their data, they placed no restrictions on the relations among the regressors; hence any hypothesized relations among the regressors are compatible with their model. For the purposes of comparing the TETRAD II output to the Rindfuss *et al.* model, the relations among the regressors are irrelevant; the only relevant part of the TETRAD II output was the part we displayed in CPS.

In CPS, we suggested a number of ways of evaluating whether or not the output should be trusted. We suggested running the program at a number of different significance levels and seeing if the output was stable; if the program was run under the assumption of no latent confounders, but the output contained double-headed arrows, it should be re-run allowing for latent confounders; and obviously, one should also ask whether the linearity assumptions made by the program are plausible. The substantive conclusions of the part of the model that we displayed in CPS pass all of these tests: DADSOCC has no direct influence on AGE, FECUND has no direct effect of ED, and ED causes AGE but not the other way around when the algorithm was run at significance levels 0.01, 0.05, 0.1, and 0.15, both assuming there were no latent confounders, and allowing for the possibility of latent confounders. Moreover, because none of the arrows was from continuous to discrete variables, the relationships could at least possibly be linear (although there is a reasonable worry about the normality assumed by some of the statistical tests).

One could, entirely irrelevantly to the point of the example, and against common sense and the advice we give both in the book and in the program manual (Scheines *et al.* [1994]), run the whole of the Rindfuss data through TETRAD II and interpret the relations among them as a causal claim we are obliged to make. That is what Humphreys and Freedman do. A large part of the model output among the regressors fails all three of the tests listed above. The model contains double-headed arrows, both the adjacencies and the orientations differ at different significance levels, and since some of the arrows are directed into discrete variables from continuous variables, the linearity assumptions have to be incorrect. The program itself indicates that the

assumptions under which it produces reliable results are false; this is as good a result as one could reasonably ask for from a program.

8. CPS contains a review of debates from the 1950s, 1960s, and 1970s over smoking and lung cancer, intended to illustrate the role of misunderstandings of the relations of causation and probability among statisticians, on the one side, and epidemiologists, on the other. We criticized the arguments historically given by both sides—epidemiologists convinced that smoking causes cancer, and statisticians convinced that no one could know.

Humphreys and Freedman denounce us for having the temerity to criticize both statisticians and epidemiologists in the debate over smoking. (pp. 118–19) They offer not a single substantive point on which they disagree with our account. They denounce us for saying that we appear not to believe the epidemiological evidence that smoking causes lung cancer (p. 118). However, we never said the *evidence* did not support the conclusion; we said the *arguments* offered did not support the conclusion.

9. One empirical question is how sensitive the reliability of the algorithms is to small violations of the assumptions, including the distributional assumptions. According to Humphreys and Freedman,

> the SGS algorithms must depend quite sensitively on the data and even on the underlying distribution: tiny changes in the circumstances of the problem have big impacts on causal inferences (p. 117).

This is a false generalization. The sensitivity of the algorithms to the distributional assumptions depends upon the sample size. At small to medium sample size, minor violations of the distributional assumptions sometimes have no effect at all on the output. It is true that in the large sample limit, the algorithms are very sensitive to violations of distributional assumptions, but of course Humphreys and Freedman have already indicated their lack of interest in the large sample limit. There is an issue here, but it is a research issue that involves work: how do the accuracies of predictions obtained from a model depend on small violations of the assumptions?

A book should fairly be judged by what it prompts as well as what it contains. Humphreys and Freedman say nothing about that, so we shall. Left unsolved was a question about making explicit all ordering information about the direction of causation implicit in a probability distribution, assuming the Markov and Faithfulness conditions and no unrecorded common causes. That question has since been solved, by Meek [1995] (and independently by Andersson *et al.* [1995] and Chickering [1995]). Meek [1995] also solved the analogous question for any prior restriction on the orientations. The book left open a conjecture about conditional independence in linear, simultaneous equation models of feedback systems, since proved correct independently by Spirtes [1995] and by Koster [1996], and generalized to a class of non-linear

feedback systems by Pearl and Dechter [1996]. These results raise questions about characterizing the statistical indistinguishability of such 'non-recursive' models, and about whether the indistinguishability classes are feasibly computable from conditional independencies, both questions since answered positively by Richardson [1996a and 1996b]. Finally, one of the standard objections to causal inference from uncontrolled samples is that because of missing values, sample convenience, and other factors, there may be *sample selection bias* that produces dependencies among variables due neither to a direct cause nor to an unrecorded common cause. Spirtes, Meek, and Richardson showed that under a modified interpretation, one of the algorithms described in the book, FCI, is correct even with sample selection bias and latent variables (Spirtes *et al.* [1995]). The graphical representation and calculation of hypothetical interventions introduced in CPS have been greatly extended in a series of articles including Pearl [1995], and Pearl and Galles [1995].

*Department of Philosophy*
*Carnegie-Mellon University*
*Pittsburg, PA 15213*
*USA*

## References

Andersson, S., Madigan, D., and Perlman, M. [1995]: *A Characterization of Markov Equivalence Classes for Acyclic Digraphs*, Technical Report 287, Department of Statistics, University of Washington.

Chickering, D. [1995]: 'A Transformational Characterization of Equivalent Bayesian Network Structures', *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, P. Besnard and S. Hanks (*eds*), San Mateo, CA, Morgan Kaufmann Publishers, Inc.

Chickering, D., Geiger, D., and Heckerman, D. [1994]: 'Learning Bayesian Networks: Search Methods and Experimental Results', *Preliminary Papers of the Fifth International Workshop on Artificial Intelligence and Statistics*.

Cooper, G. and Herskovits, E. [1992]: 'A Bayesian Method for the Induction of Probabilistic Networks from Data', *Machine Learning*, **9**, pp. 309–47.

Fisher, R. [1960]: *The Design of Experiments*, 7th edn, New York, Hafner Publishing Company.

Freedman, D. [1994]: 'From Association to Causation Via Regression', Technical Report 408, Statistics Department, University of California, Berkeley; also to appear in Vaughan McKim (*ed.*), *Causality in Crisis: Statistical Methods and the Search for Causal Knowledge in the Social Sciences*, University of Notre Dame Press.

Freedman, D. [forthcoming]: 'Reply to Spirtes and Scheines', in Vaughan McKim (*ed.*), *Causality in Crisis: Statistical Methods and the Search for Causal Knowledge in the Social Sciences*, University of Notre Dame Press.

Humphreys, P. [1989]: *The Chances of Explanation: Causal Explanation in the Social, Medical, and Physical Sciences*, Princeton, NJ, Princeton University Press.

Humphreys, P. and Freedman, D. [1996]: 'The Grand Leap', *British Journal for the Philosophy of Science*, **47**, pp. 113–23.

Koster, J. [1996]: 'Markov Properties of Non-Recursive Causal Models', *Annals of Statistics*, **24**, pp. 2148–2177.

Meek, C. [1995]: 'Causal Inference and Causal Explanation with Background Knowledge', *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, P. Besnard and S. Hanks (*eds*), San Mateo, CA, Morgan Kaufmann Publishers, Inc., pp. 403–10.

Pearl, J. [1995]: 'Causal Diagrams for Empirical Research', *Biometrika*, **82**, pp. 669–710.

Pearl, J. and Galles, D. [1995]: 'Testing Identifiability of Causal Effects', *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, P. Besnard and S. Hanks (*eds*), San Mateo, CA, Morgan Kaufmann Publishers, Inc., pp. 185–95.

Pearl, J. and Dechter, R. [1996]: *Identifying Independencies in Causal Graphs with Feedback*, Technical Report R-234, Cognitive Systems Laboratory, University of California at Los Angeles.

Richardson, T. [1996a]: 'A Discovery Algorithm for Directed Cyclic Graphs', *Proceedings of the Twelfth Conference on Uncertainty in Artificial Intelligence*, E. Horvitz and F. Jensen (*eds*), San Mateo, CA, Morgan Kaufmann Publishers, Inc.

Richardson, T. [1996b]: 'A Polynomial Time Algorithm for Deciding Markov Equivalence of Directed Cyclic Graphical Models', *Proceedings of the Twelfth Conference on Uncertainty in Artificial Intelligence*, E. Horvits and F. Jensen (*eds*), San Mateo, CA Morgan Kaufmann Publishers, Inc.

Rindfuss, R., Bumpass, L., and St. John, C. [1980]: 'Education and Fertility: Implications for the Roles Women Occupy', *American Sociological Review*, **45**, 431–47.

Scheines, R., Spirtes, P., Glymour, C., and Meek, C. [1994]: *TETRAD II: Tools For Causal Modeling*, Hillsdale, NJ, Lawrence Erlbaum.

Spirtes, P. [1995] 'Directed Cyclic Graphical Representation of Feedback Models', *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, Philippe Besnard and Steve Hanks (*eds*), San Mateo, CA, Morgan Kaufmann Publishers, Inc.

Spirtes. P, and Scheines, R. [forthcoming] 'Reply to Freedman', in Vaughan McKim (*ed.*), *Causality in Crisis: Statistical Methods and the Search for Causal Knowledge in the Social Sciences*, University of Notre Dame Press.

Spirtes, P., Glymour, C., and Scheines, R. [1993]: *Causation, Prediction, and Search*, Lecture Notes in Statistics 81, New York, Springer-Verlag.

Spirtes, P., Meek, C., and Richardson, T. [1995]: 'Causal Inference in the Presence of Latent Variables and Selection Bias', *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, Philippe Besnard and Steve Hanks (*eds*), San Mateo, CA, Morgan Kaufmann Publishers, Inc.

Verma, T. and Pearl, J. [1990]: 'Equivalence and Synthesis of Causal Models', *Proceedings of the Sixth Conference on Uncertainty in AI*, Association for Uncertainty in AI, Inc., Mountain View, CA.