Causation, Prediction, and Accommodation

Malcolm R. Forster
mforster@facstaff.wisc.edu

December 26, 1997

ABSTRACT:

Causal inference is commonly viewed in two steps: (1) Represent the empirical data in terms of a probability distribution. (2) Draw causal conclusions from the conditional independencies exhibited in that distribution. I challenge this reconstruction by arguing that the empirical data are often better partitioned into different domains and represented by a separate probability distribution within each domain. For then their similarities and the differences provide a wealth of relevant causal information. Computer simulations confirm this hunch, and the results are explained in terms of a distinction between prediction and accommodation, and William Whewell's consilience of inductions. If the diagnosis is correct, then the standard notion of the empirical distinguishability, or equivalence, of causal models needs revision, and the idea that cause can be defined in terms of probability is far more plausible than before.

1 Cause and Correlation

Someone knocks on your door selling subscriptions to the *Wall Street Journal*. "Did you know," the sales pitch begins, "that the average reader of the *Wall Street Journal* earns more than $70,000 a year?" You have just been told that there is a positive correlation between "*Wall Street Journal* reading and income level, and you have no reason to dispute the premise. But should you conclude from the premise that your subscribing to the *Wall Street Journal* will increase the chances of increasing your income to over $70,000 (supposing that you are a philosopher)? Anyone with an ounce of common sense knows that the causal claim does not follow from the stated correlation.[1]

Causal inference is not easy, and there are many complications besides the one illustrated in this example. However, there has been much progress in recent years on the problem of causal inference, and even automated causal reasoning done by computers (see Korb and Wallace (1997) and Spirtes *et al* (1997) for brief surveys). The purpose of my article is to argue that this recent work is based on an incomplete picture of causal inference.

Think of causal inference as having two parts. First, one uses the statistical data to make inferences about probabilities. This involves standard forms of statistical inference, and here there are disputes about which statistical methodology is best. For example, Spirtes *et al* (1993) favor a more classical approach, whereas Korb and Wallace (1997) prefer a Bayesian

[1] I believe that I first read an example like this in Cartwright (1983).

methodology. However, all sides agree that the hard and interesting part of *causal* inference arises in the second step, in which causal conclusions are drawn from the probabilistic facts. They also agree that this second step will, at best, lead to a class of causal models that are indistinguishable (Spirtes *et al* (1993)), empirically equivalent (Verma and Pearl (1991)), or observationally equivalent (Hoover (1993)). That is, the probabilistic facts must under-determine the causal facts. This under-determination shows that the concept of cause presupposed in standard causal modeling cannot be defined in probabilistic terms. Cause does not reduce to probability, in other words. If the received view of causal inference is correct, then this conclusion is correct. But the received view is not correct.

I would be committing the fallacy of denying the antecedent if I were to conclude that cause does reduce to probability. But I can at least conclude that reductionism is, once more, an open question. Reductionist views are still criticized in the recent literature (Hausman, forthcoming), and there has been recent dispute between Papineau (1989, 1991, 1993) and Irzik (1996). So, my main thesis, if correct, promises re-invigorate the reductionist side of the debate.

I agree that causes do not reduce to correlations, partial correlations, or conditional dependencies. However, if one includes information about how correlations *change* from one situation to the next, or more exactly, how correlations do not change, then a reductionist view of causation is more plausible. There has been an active discussion of this idea of invariance in philosophical literature over the years. It was something I called cross-situational invariance in Forster (1984), which was taken up and discussed in Hooker (1987). The notions of resilience (Skyrms 1980), homogeneity (Salmon 1971), and robustness (Redhead 1989) are related, although a closer approximation is found in Arntzenius (1997), Harper (1989), Hoover (1994), Simon (1953), Sober (1994), and Woodward (1993, 1997, 1998). An early development of the epistemological idea comes in the form of Whewell's famous notion of the consilience of inductions *circa* 1840 (see Butts (1989) for a good collection of Whewell's writings in the philosophy of science). Whewell values the agreement of independent measurements of theoretical quantities as one of the most persuasive kinds of confirmation in the history of science. Forster (1988) is an application of Whewell's consilience of inductions to a problem Cartwright (1983) raises about the under-determination of component forces in Newtonian mechanics (a violation of the facticity requirement, as she put it).

Despite this wide ranging discussion in the philosophical literature, the idea is not built into the design of any of the standard methods of automated reasoning in existence today, as far as I am aware.

Here is a simple example to motivate the relevance of the idea in causal reasoning (Arntzenius (1997)). At a time when there were relatively few smokers, the percentage of lung cancer victims who were smokers was practically zero. Yet nowadays, especially in countries with higher numbers of smokers, that percentage will be much higher. That is, the chance that someone smokes given that they have lung cancer depends on the base rate of smokers. Yet, the forwards probability of lung cancer given smoking will be relatively stable. This is just as the hypothesis that smoking causes lung cancer predicts. But note that this prediction goes beyond the simple prediction that there is a positive correlation between smoking and lung cancer. That is why a reductionist should not expect that cause reduces to correlational facts alone.

Spirtes *et al* (1993) recognize that correlations alone do not uniquely determine a causal hypothesis. The same point is clearly stated by the originator of causal modeling (Wright (1923, p.254) in response to critics: "The method of path coefficients does not furnish general formulae for deducing causal relations from knowledge of correlations and has never claimed to do so." But instead of concluding that there must be other kinds of *empirical* information, Spirtes *et al* (1993) authors embrace the non-reductionist conclusion.[2] Their view is that this is the best we can do. The plausibility of their position rests rather heavily on the idealization, mentioned earlier, that the data is correctly and completely represented in terms of a single probability distribution of the space of possible events. But if there is other empirical information available, then this is not the best we can do.

My main thesis is that information about the variability of correlations is most often available and it is causally relevant. Therefore, many standard approaches to causal inference, including Spirtes *et al* (1993), are incomplete. The hard problem is to say what invariance is and how it is relevant to causal inference. That is the task I will begin in this paper.

The paper is organized as follows. Section 2 explains the recent debate about the reduction of cause to probability, and argues that the well known distinction between the prediction and the accommodation of evidence is relevant to this debate. Here I begin to explain how the added element of invariance ties into these issues. Section 3 describes a theorem (in Verma and Pearl (1991)) about the empirical indistinguishability of causal models found in the literature on automated causal inference. In section 4, I examine what causal models say about probabilities and the invariance of probabilities, and this is extended in sections 5 and 6. The crux of this paper comes in section 7, where I argue that the correct understanding of the equivalence of causal models is quite different from what Verma and Pearl (1991) assume. Yet, the question about the *empirical* distinguishability of causal models is not really resolved until section 8, where I use computer simulations to demonstrate how Whewell's consilience of inductions adds new information about the invariance of correlations. Finally, this is tied into the earlier discussion of prediction and accommodation.

2 Prediction versus Accommodation

Cause is not the same as correlation. For example, suppose that we find that there is a higher frequency of heart disease amongst coffee drinkers than amongst the rest of the population. Should we conclude that coffee drinking causes heart disease? Not on the basis of this evidence, for there is an alternative explanation of the correlation: coffee drinkers tend to smoke more frequently than those who do not drink coffee, and smoking causes heart disease. If this explanation is correct, then giving up coffee will not help prevent heart disease (whether you are a smoker or not). You should give up smoking instead.

If there are just two variables, coffee drinking c, and heart disease h, then the correlational facts do not determine that c causes h because other causal hypotheses predict the existence of the correlation. Therefore, cause does not reduce to correlation. The standard reply to this argument is to say that we need to consider background variables. If we include the correlations of smoking s with c and h, then there may be a correlational asymmetry that resolves the ambiguity in the causal explanations. For example, suppose that the same studies that show a correlation between coffee drinking and heart disease also show that the correlation disappears once the smokers are separated from the nonsmokers. Amongst smokers, coffee drinking is not correlated with heart disease, and amongst non-smokers, coffee drinking is also uncorrelated with heart disease. (That this is consistent with there being a correlation between coffee drinking and heart disease in the population as a whole is know as Simpson's paradox.) On the other hand, the correlation between smoking and heart disease does not disappear in the same way. Smoking increases the chance of heart disease amongst coffee drinkers and non drinkers alike. So, there is a correlational asymmetry between smoking and coffee drinking with respect to heart disease, which may

[2] It is unclear to me whether Wright draws the same conclusion, for he also says (p. 241) that he "accepts the viewpoint that our conceptions of causation are based merely on experience." And later (p. 252): "The formulation of hypotheses is emphatically the business of one who is thoroughly familiar with the realities of the case." If he is a reductionist, then he certainly fails to give a clear account of the "experience" or the "realities of the case" on which the concept of causation is based.

ground the causal asymmetry between coffee drinking and smoking in their relationship to heart disease.

Philosophers commonly describe this by saying that smoking (or non-smoking) *screens-off* coffee drinking from heart disease, while coffee drinking does not screen-off smoking from heart disease. Statisticians describe the same facts in terms of partial correlations, and computer scientists and statisticians refer to conditional independencies. Facts equivalently described in terms of the presence or absence of partial correlations, screening-off relations, or conditional independencies, all count as correlational facts for the purposes of this paper.

Irzik (1996) claims that this asymmetry between *c* and *s* in relationship to *h* does not resolve the ambiguity in the direction of the causal arrow from *s* to *h*. For the observed screening-off relations in our simple example is *compatible* with a quite different causal model (see Figure 1); one in which present smoking is an effect of having heart disease in the future. One reason we tend to dismiss the model in Figure 1 is that it postulates backward causation. But reductionists like Papineau, and myself, decline to use temporal restrictions in their analysis, for then it is no longer a reduction to probabilistic facts alone. So, the existence of this alternative explanation is a real problem for anyone who thinks that the probabilistic facts uniquely determine the causal structure.

At this point, it is possible for a reductionist to invoke the same defense as before. Perhaps if we include still further background variables, a further correlational asymmetry will emerge that will resolve the current remaining causal ambiguity. However, it is implausible that *all* ambiguities will ever be resolved because we introduce new ambiguities every time we add background variables. In fact, the theorems of Verma and Pearl (1991) and Spirtes *et al* (1993) support this view. This is somewhat disconcerting from an inferential point of view. For it seems that we cannot unambiguously infer that smoking causes heart disease even if we know every correlational fact there is.

For that reason, I wish to explore the possibility that other kinds of information are relevant to causal questions. One valuable insight is that there is a sense in which Irzik's alternative model does not *explain* or predict the screening-off relation, since the model in Figure 1 does not require it. The model merely *accommodates* the screening-off relation in the sense that it is consistent with it. In contrast, the model that we all believe predicts that the screening-off relation *must* occur.

A famous example of the difference between prediction and accommodation arose in the way Ptolemy and Copernicus accounted for various regularities of planetary motion. One noticeable phenomenon is that, while planets generally wander against the fixed stars in the same direction, they sometimes move backwards for a while, and then forwards again. The backwards motion is called retrograde motion. But even more interestingly, the retrograde motion of the outer planets occurs when and only when the Earth is between the Sun and the retrograding planet. That is, the outer planets retrograde when and only when in *opposition* to the Sun. Copernicus claimed to have *explained* this phenomenon as a *necessary* consequence of his model. His explanation went like this: All the planets revolve in the same direction and the inner planets revolve more quickly, and therefore overtake the outer planets periodically. When that happens, the planet appear to move backwards against the fixed stars. But this 'overtaking' can only occur when both planets are on the same side of the Sun because the Sun is in the middle of both orbits. In Ptolemy Earth-centered theory, on the other hand, the sun was not at the center of the planetary orbits,

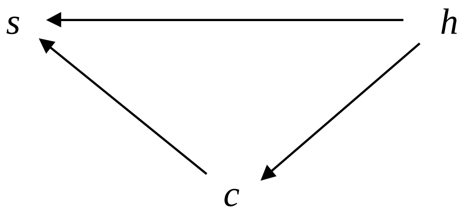

Figure 1: Irzik's alternative model. *s* = smoking, *c* = coffee drinking, *h* = heart disease

so there was no such necessary consequence. However, Ptolemy's geocentric theory could *allow* that the retrograde motion of planets occurs only in opposition to Sun. Ptolemy accommodated the observed phenomenon, while Copernicus *predicted* (or postdicted) the phenomenon.

This contrast is not an artifact of this example. If we precede further in the history of planetary astronomy, we see that Newton's theory succeeded in predicting (or postdicting) that the inner planets will revolve more quickly, while Copernicus was only able to accommodate this fact. Examples of the same distinction abound throughout the history of science.

Irzik's example exhibits the same distinction. There is a phenomenon—a screening-off relation—exhibited in the statistical data. It is predicted by the hypothesis that smoking causes heart disease, but it is not predicted by the alternative causal model in Figure 1. That model is only able to *accommodate* the screening-off. The smoking hypothesis is 'Copernican', while the alternative hypothesis is 'Ptolemaic'. One model predicts observed screening-off while the other model does not. So, there is a straightforward empirical reason for favoring the 'Copernican' model.

However, the situation is not quite that simple. The first complication is that the very best models in science do not, and should not, *predict* all observational facts. There is always some accommodation needed. For example, Newton's theory of gravitation did not predict the masses of the Sun and the planets, or their initial positions and velocities. Those aspects of the phenomena are merely accommodated. Once accommodation is complete, the theory makes precise predictions of everything else, but that is perfectly analogous with Ptolemy's model.

But what exactly is the methodological lesson here? It could be that a theory should predict as much of the phenomena as possible, and the model that successfully predicts the most is favored. Or it may be that there is some prior analysis of which aspects should be predicted and which should be accommodated. I believe that the correct answer is a mixture of both.

In an interesting paper, Bogen and Woodward (1988) develop the idea that there is a two step procedure involved in comparing any theory with the observational data. The first step is to infer the *phenomena* from the observational data. For instance, first establish an empirical generalization, often called 'effects' in physics (the Zeeman effect, the Hall effect, the photo-electric effect, and so on). Then look for explanations of those phenomena. However, even the best theories in the history of science do not succeed in predicting every aspect of the known phenomena. For example, the fact that all planets revolve around the Sun in the same direction, though predicted by Descartes' vortex theory, was merely accommodated by Newton's theory of gravitation.

Which aspects of the phenomena should a theory predict, and which parts should it merely accommodate? There appears to be no *a priori* answer to this question, for it is a question about how we should compare theories. However, there is a second question we could ask: Which aspects of the phenomena *can* a theory predict, and which parts can it not predict? To answer this question, we need to examine the theory closely.

This two-step process bears a striking resemblance to our previous division of causal inference into two steps. The first is an inference to probabilistic facts, and the second is an inference from probabilities to causes. The inferred probabilistic facts play the role of phenomena in the Bogen and Woodward scheme. But which aspects of the phenomena can a causal model predict, and which parts should a model merely accommodate? We know that a causal model can predict conditional independencies, and I think it should. But whether this is *all* that a causal model can predict depends on the causal models, which is why I plan to examine them carefully in later sections.

For now, my point is that the standard analyses of causal inference commit themselves to the as-

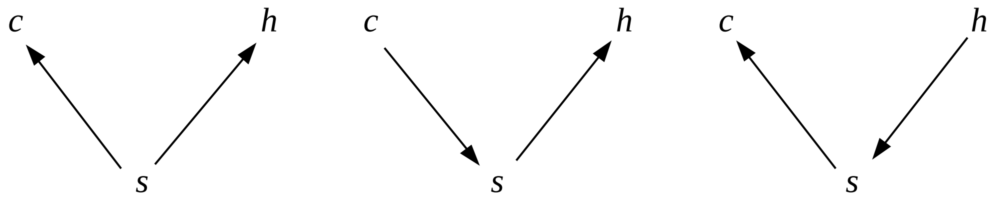

Figure 2: All three causal models predict the same conditional independencies.

sumption that a model should predict all the conditional independencies exhibited in a single (statistically inferred) probability distribution. From this principle, we may eliminate Irzik's example on the basis of the empirical data. However, there are other causal models that cannot be eliminated. In Figure 2, the model on the far left, in which smoking is a common cause of coffee drinking and heart disease, is the one we believe. But the two models to the right of it predict exactly the same conditional independencies. Therefore, the correlational facts underdetermination the causal facts. In the next section, I will examine the argument in more detail.

3 The Alleged Indistinguishability of Causal Models

Let me begin by describing causal inference according to Spirtes *et al* (1993).[3] First, some preliminary definitions. First, we need to introduce the terms "parent", ancestor", and "descendent". A variable *a* is a *parent* of a variable *b* if and only if there is an arrow from *a* to *b*. The parent of a parent, or the parent of a parent of a parent, and so on, are called *ancestors*. A parent, in other words, is an immediate ancestor. Finally, *b* is a *descendent* of *a* if and only if there is a causal path from *a* to *b* that moves along arrows in the forward direction. A system *S* of variables is *causally sufficient* if and only if for every pair of variables *a*, *b* in *S*, if there is a variable *c* from which there is a causal path to *a* and also a causal path to *b*, and the two paths do not intersect each other except at *c*, then *c* is in *S* as well. That is, for every pair in *S*, every *common* ancestor is also in *S*. A causally sufficient system is one in which all common causes are included. There are no "latent" common causes, in other words.

The next step is to say what conditional independencies are predicted by a sufficient causal model without loops (such models are called *acyclic*).

Markov Condition: In a causally sufficient system described by an acyclic causal model, conditional on any set of values of all of its parents, every variable is independent (in probability) of the set of its non-parent non-descendents.

In Figure 2, assuming that {*c*, *s*, *h*} is causally sufficient, the Markov Condition implies exactly the same conditional dependency; namely that *c* and *h* are independent conditional on any fixed value of *s* (smoking or non-smoking). By the same token, there are no conditional dependencies predicted by Irzik's model in Figure 1, despite the fact that the model can accommodate the same conditional independence. If conditional dependencies are the only phenomena that causal models predict, then it follows that the three models in Figure 2 are empirically indistinguishable. Of course, I plan to dispute the antecedent of this conditional.

In an interesting paper, Verma and Pearl (1991) claim to give a complete characterization of when two causally sufficient models are equivalent according to a technical definition that I will not repeat. They interpret this definition to capture the idea that "Two causal models are equivalent if there is no experiment which could distinguish one from the other." (Verma and Pearl (1991), p. 255.) The whole point of my paper is to show that their definition does not capture the idea of empirical distinguishability at all, so I will refer to their technical notion of equivalence as *VP-equivalence*. Their no-

[3] My exposition is borrowed from Glymour (unpublished).

tion is correctly interpreted as saying that two causal models are VP-equivalent if there is no experiment which could distinguish one from the other *on the basis of conditional independencies alone*. Here is one of the theorems they prove.

Theorem 1: Two causal models are VP-equivalent if and only they have the same links and the same uncoupled head-to-head chains. To say that two variables are *linked* is that say that they are connected by an arrow. $a \rightarrow c \leftarrow b$ is a *head-to-head chain*. It is *uncoupled* if and only if a and b are unlinked.

So, all the models in Figure 2 are VP-equivalent by this theorem because they have the same links, and none of them have head-to-head chains. However, they are not VP-equivalent to Irzik's model in Figure 1 because it has an extra link (although it also has no *uncoupled* head-to-head variable). On the other hand, the model in Figure 3 is not VP-equivalent to any of these models, because although it has the same links as the models in Figure 2, it has an uncoupled head-to-head variable.

Despite the fact that Irzik's model is empirically distinguish*able* from the three models in Figure 2 (my definition and Verma and Pearl's share this consequence), it does not follow that the model is in fact distinguish*ed* from them. Suppose that Irzik's model were true, and per chance, c and h come out to be independent conditional on s. Then any causal inference based on conditional independencies would eliminate the true model from contention because it did not predict the phenomenon. Spirtes *et al* (1993) do not see this as a problem because such mistakes will be rare. They proceed on the assumption that none of the conditional independencies exhibited in the data arise from such accidental circumstances, which they call the Faithfulness Condition. I agree with them that it is a methodologically respectable assumption to make. Every scientific inference is equally prone to the same kind of mistake. It could have been possible that Ptolemy's geocentric theory was true, and that the fact that retrograde motion of the outer planets always occurred in opposition to the Sun was a mere coincidence. Such risks are the price of business in science. On my view, the Faithfulness Condition is founded on a healthy respect for the difference between prediction and accommodation.

The Markov Condition is more problematic. The problems come in two forms. The first problem is about whether it is always true in a causally sufficient set of variables, while the second problem concerns the very existence of that precondition—that a set of variables is causally sufficient. I will discuss these in turn.

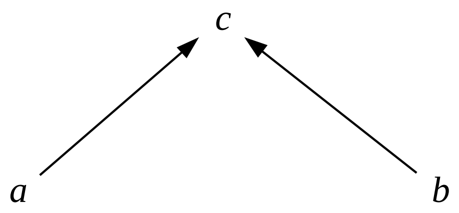

Figure 3: *a* and *b* each cause *c*.

Consider the case in which a and b are causes of c (Figure 3). The Markov Condition implies that a and b are probabilistically independent. Suppose that a and b are probabilistically dependent, which is to imply that they are correlated. This finding is consistent with the causal model in Figure 3 if and only if there is a common cause variable not shown in the diagram. However, there are several ways in which uncaused correlations may arise.[4] For one, there are strange correlations in quantum mechanics for which there is no common cause. In response, Spirtes *et al* (1993, p. 64) state that "In our view the apparent failure of the Causal Markov Condition in some quantum mechanical experiments is insufficient reason to abandon it in other contexts." Their response is quite reasonable from a practical point of view.

However, there are ways in which causally unrelated variables may become entangled as a pure artifact of the way the data is pooled together, which may happen in *any* example. Suppose that there are

[4] The fact that some correlations need not arise as the result of a common cause is well documented in the philosophical literature under the title "counterexamples to Reichenbach's principle of common cause." See Arntzenius (1993) for a list of these counterexamples, with back references to the philosophical literature on the subject.

two sets of experimental data, such that *a* and *b* are independent within each of them, but that each variable falls in a different range of values in each case. For example, *a* and *b* may take on a low range of values in the first data set, but high values in the second data set. When we combine the two data sets into one, then *a* and *b* are correlated. You can understand why this happens by thinking of correlation as an indicator relation: *a* is correlated with *b* if and only if one can make a better prediction about the value of *b* by knowing the value of *a* than without knowing it. The variables are correlated in the pooled data because information about *a* tells you which data set it's from, which in turn gives you information about the value of *b*. Automated causal inference looks at the conditional independencies in a single probability distribution; presumably the one that matches the *pooled* data. So artifacts of this kind could be a problem.

However, these doubts about the Markov condition simply compound a problem that exists. For if we need to know whether a system of variable is causally sufficient before drawing causal conclusions, then any inference from observed conditional independencies is powerless. The worst case scenario is that, as Pearson once put it (quoted by Wright (1923, p.250)), "The causes of any *individual* thing thus widen out into the unmanageable history of the universe. The causes of any finite portions of the universe lead us irresistibly to the history of the universe as a whole."[5] In fact, the situation is not as bad as that, as Verma and Pearl (1991) and Spirtes *et al* (1993) show in some detail (but also see Robins and Wasserman (1998)). Nevertheless, if we are unable to eliminate the possibility of 'latent' common causes, our causal conclusions are even more ambiguous than what is already entailed by Verma and Pearl's Theorem 1. There are severe constraints on what can be inferred from correlational facts alone, and this motivates an interest in other possible sources of empirical information.

Here is a first step in that direction. Consider the following principle:

Principle of functional autonomy: The mean value of a variable *c* conditional on values of its parents {*a*, *b*} does not dependent on the probability relations amongst the parents.

For example, the function $z = f(x, y)$ defines a mapping from (x, y) values to *z* values, which does not depend on the probabilities of *x* or *y*. The same applies to conditional probabilities. For example, the probability of *A* given *B* and *C*, viewed as a mapping from *B* and *C* to a probability value for *A*, does not depend on the probability of *B* and *C*. This is a basic fact about functional relationships, so the principle not new or controversial.

Some of the applications of this principle that have important nontrivial consequences. Suppose that a causal model asserts that the causal dependencies of *c* on *a* and *c* on *b* are the same in two different causal contexts: In the first situation *a* and *b* are causally unconnected, while in the second situation, *a* is the cause of *b* (see Figure 4). The 'composite' causal hypothesis specifies exactly what is the same and what is different in the two situations. By hypothesis, *c* has exactly the same functional dependence on *a* and *b* in both situations. Let me refer to a model that assumes that some causal relations are

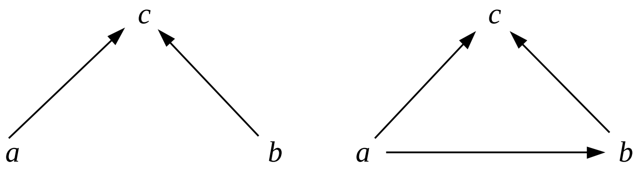

Figure 4: *c* has the same causal dependence on *a* and *b* in both situations

[5] Wright (1923, p.250) claims that this problem arises in any attempt at discerning the relative importance of heredity and environment in determining the characteristics of a single given individual. As he says, "the genetic constitutions of two guinea-pigs, chosen at random from a certain stock, undoubtedly trace back to numerous common causes." However, he adds that the problem of determining the relative importance of the *variation*, or *differences*, within a given stock can be "solved with great ease" because "in subtracting the total cause of one event from another there is an enormous cancellation of common causes."

cross-situationally invariant as a *unified* model. Causal inferences commonly rely on such models when manipulations break some, but not all, causal relationships. The Markov Condition captures the differences between such situations, but not their similarities. As far as conditional independencies go, the two situations have nothing in common (remember that the correlation between *a* and *b* marks their difference, not their similarity).[6] So, under the received view, it is not clear that one is ever any reason to favor the unified model over and a disunified model that treats the two situations having *nothing* at all in common. Or to put the point a different way, if one is able to favor the unified model, then there is a tacit appeal to evidence that goes beyond conditional independencies; namely, the invariance of the *functional* dependence of *c* on *a* and *b*.

As charity demands, I am assuming that the data pertaining to the two situations is not pooled in this case. If it pooled, then the received view faces the opposite problem: There are can be no inferred differences. Either way, something is being missed.

It appears that not only conditional independencies but also conditional dependencies are relevant to causal inference. But is this a serious concession for the received view? Isn't it the case that the received view automatically takes this information into when comparing the fit of the unified model and the disunified model? Unfortunately, this hunch is incorrect. It is not the existence of conditional dependencies that is relevant, but the fact that they do not *change* from one situation to the next. This information is not utilized in the standard methods of causal inference.

While the nature of this new empirical information is still unclear, its existence is enough to undermine the standard interpretation of the 'indistinguishability' theorems. These theorems do not establish that VP-equivalent models are *empirically* indistinguishable, as Verma and Pearl (1991, p.255) claim. I hope to drive this point home in the final sections.

4 The Modal Content of Causal Models

Here is the simplest case of two causal models that are indistinguishable by conditional independencies alone: $a \rightarrow b$ versus $a \leftarrow b$ (they have the same links and the same uncoupled head-to-head variables). I imagine that if I can show that these two models are empirically distinguishable, without extending the context to include new variables, then the point will have an impact. For the simplest case is the hardest case.

Therefore, I will consider two such models within the framework of *path analysis* (Wright 1921, 1923), which is the same framework assumed by Irzik (1996).[7] In this section and the next, I shall discuss the content of such models, and then extend the discussion to include their *empirical* content and *empirical* distinguishability in later sections.

To make the example less abstract, consider one of those dimmer switches that adjust the intensity of a ceiling light. Let *X* be the angle at which the knob is turned, and *Y* the intensity of light. In this example, the argument against the reduction of cause to probability is that the correlation between *X* and *Y* is symmetric, and cannot mark the difference between *X* causing *Y* (which is what we believe) and *Y* causing *X*. In this simple example, there are two causal models to consider:

Model 1: $Y = \alpha_0 + \alpha_1 X + U$

Model 2: $X = \beta_0 + \beta_1 Y + U'$

Model 1 is the model we use if *X* causes *Y* and Model 2 is the one to use if *Y* causes *X*. In each case, the

[6] As charity demands, I am assuming that the data pertaining to the two situations is not pooled in this case. If it pooled, then the received view faces the opposite problem: There are can be no inferred differences. Either way, the received view is in trouble.

[7] I am assuming that *causal* models, so-called, do genuinely capture the notion of cause. Humphreys and Freedman (1997) complain that such assumptions are unjustified. But if causal models do not capture the notion of cause, then they don't say what does. Whatever the answer, the concept that *is* captured by causal modeling is a rich, powerful, and interesting, notion worthy of discussion.

variable on the left is called the dependent variable, and the other variables are called independent variables (the terminology is a little confusing because 'independent' here has nothing to do with probabilistic independence).

Note that the model is written here for a single instance, so X and Y refer to 'token' events.[8] In the instance in question, X, Y and U will have particular values. The U term is often referred to as the *residue* term. The idea is that an equation stripped of the residue term, like $Y = \alpha_0 + \alpha_1 X$, represents the 'signal', the trend, or the regularity between Y and X, while U is the *noise*, or *error*. ('Noise' would be the most appropriate term for this paper, but it is not standard. 'Error' has unwanted connotations, so I will use the term 'residue'.) Alternatively, the equation $Y = \alpha_0 + \alpha_1 X$ may be thought of as describing the *mean curve* because standard assumptions made about the residue term imply that, for any fixed value of X, the mean value of Y is equal to $\alpha_0 + \alpha_1 X$.

The model claims that the value of Y is a certain function of the values of X and U. If there are many events under consideration, as is usual, then the same equation is applied to them as well, *with the same parameter values*. The αs and the βs are constants. However, it is wrong to think of exactly the same variables, X, Y, and the U, as applying to new set of events. They should be subscripted, to indicate that different instances are involved. I will say more about this in a moment.

Let me concentrate on Model 1, so that Y is the dependent variable and X is the independent variable. Everything to be said will extend to Model 2 by taking X to be the dependent variable and Y the independent variable. The term U_Y is the residue, which allows the model to cover the situation in which X alone does not determine the value of Y exactly, but only approximately or probabilistically. Probability enters the model via probability assumptions made the residue. In particular, U is assigned a probability distribution for every value of the independent variable X. The standard assumption, which I adopt, is that the mean value of U is zero in each case. This is why the equation with the residue term stripped away is the mean curve of Y on X. There are no constraints on the variances or other features of the residues.

By introducing these assumptions, we automatically constrain the joint probabilities of the all the variables because they are connected together in a single equation. More specifically, we know the probability distribution of the dependent variable conditional on any value of the independent variable. Note that this applies even when Y is a nonlinear function of X. Also note that we have introduced an asymmetry between X and Y, because we determine the distribution of Y on X, but not the distribution of X on Y, nor the distribution of any of the independent variables. (This is another application of the principle of functional autonomy introduced in the previous section.) This asymmetry is represented by directed graphs like those in Figure 5.

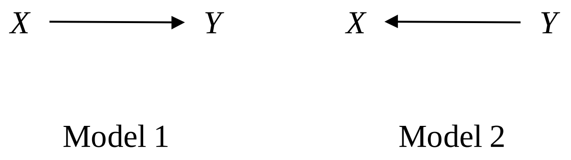

Figure 5

One sometimes sees the argument that Model 1 and Model 2 are equivalent because the equation of Model 1 can be transformed into the equation of Model 2 in the following way:

$$X = -\frac{\alpha_0}{\alpha_1} + \frac{1}{\alpha_1} Y - \frac{1}{\alpha_1} U \ .$$

However, this is incorrect. It is true that the parameters in this transformed equation are still constants, and that the (unconditional) mean of the residue term is zero, but the residue term does not have mean zero conditional on any value of Y. The point is that the asymmetry between X and Y in Model 1 does not arise from the way that the equation is

[8] Hitchcock (1995) discusses the distinction between 'token' and 'type' causation, or 'singular' and 'general' causation in a way that nicely complements what I am going saying.

written, but from the probabilistic assumptions made about the residue.

Therefore, the content of Model 1 is not determined by the equation $Y = \alpha_0 + \alpha_1 X + U_Y$ alone. Certainly, the equation plays a crucial role, but *by itself* it is merely a definition of U_Y. Many people seem surprised by this claim, so let me argue for it. The point is clearest when we think of the equations in terms of standard regression analysis. Let $\overline{Y} = \alpha_0 + \alpha_1 X$ be some curve in the *X*-*Y* plane. The residue, is *defined* as $Y - \overline{Y}$, where *Y* is the true value of *Y* (the one that is observed, say) and the sum of these residues squared is called the sum of squared errors, or the sum of squared deviations. This definition applies universally to all such examples. We therefore write $Y = \alpha_0 + \alpha_1 X + U$, where $U \equiv Y - \overline{Y}$.

U is a random variable because it is a function of *X* and *Y*, which are random variables (a random variable is just a variable whose values are assigned a probability). Although the residue *may* be explained as arising from the action of other causes, that is not necessarily true. In an indeterministic world such as ours, if quantum mechanics is a complete description of reality, there may be no such causes. Yet the residue is still well defined, and causal models still apply. So there are no causal assumptions built into the *definition* of *U*. For example, suppose we are measuring the decay rate of a piece of radioactive uranium over time. There is nothing to prevent us from applying a standard regression analysis to this phenomenon, despite the fact that the random deviations of the actual decay numbers from the mean (exponential) curve is not explained as arising from other causes. Thus, contrary to what Irzik claims (1996, p. 261), it is not essential that the residue *U* refer to an event. It is not a catchall for 'other causes'.

Thus far, I have described the content of Model 1 in terms of its probabilistic consequences. It is important that this does not, immediately, describe the *empirical* consequences of the model. For probabilities, as I understand them, are defined over a space of *possible* events, and not a set of actual events. They are not defined in terms of the frequency of instances in a population of similar event. The term 'modal' refers to the notion of 'possibility', so I describe the content of Model 1 as *modal*. It does have empirical content when (but only when) auxiliary assumptions are added to the model, but the topic of this paper requires me to separate these very carefully. So, let me continue to describe the content of Model 1.

As an illustration, suppose I am about to turn the dimmer switch 90°, and wish to consider its effect on the light intensity. The event will be *X* = 90° followed by the event *Y* = 30 lumens, say. But a probability function applies not just to these individual events, but to every pair of possible events (*X* = *x*, *Y* = *y*), *for all* values of *x* and *y*. The set of all possible event pairs is called the *event space*. A probability over an event space entails much more than the probability of the occurrence of the actual events *X* = 90° and *Y* = 30 lumens. To put the point more dramatically, though less accurately, the probabilistic hypothesis supports counterfactual conditionals.[9] It entails the probability of the effect given *counterfactual* events like *X* = 45°. A probability function is strongly modal (it refers to possibilities) and the reduction of cause to probability is therefore more plausible than might be supposed at first glance.

Consider Model 1 (or Model 2) applied to a single actual pair of event; say *X* = 90° and *Y* = 30 lumens. The same equation also applies to the space of counterfactual events of the same type (with the same variable values) as well as to other event types (with other variable values). That is the sense in which *X* and *Y* are random *variables*—they range over a space of *possible* events. On the other hand, the αs and the βs are constants, because they have the same value no matter which possible events are considered.

[9] The semantics of counterfactual conditionals is not only modal, but also makes use of a similarity metric over possible worlds, or something equivalent, while probabilistic assertions do not. So an appeal to counterfactuals would be more problematic.

Given the importance of this point to what follows, allow me to speculate about why the modal character of probabilities is often overlooked. The first is the common urn model of probability. If we randomly choose a marble from an urn containing 9 white marbles and one black one, we *calculate* the probability of drawing the black marble from the proportion of *actual* black marbles in the urn. But the probability itself is not the proportion of black marbles in the urn. It pertains to an *event* space, which in this case is the space of all *possible* drawings from the urn in a particular instance.

Sometimes the modal nature of probability is obscured when the conclusion of a statistical inference is an assertion about proportions in a population of actual events. The interpretation of polling results is an example. Here we want to know the proportion of actual American voters would vote Democrat if the election were held today from a poll. These proportions are not probabilities over a space of possible events, but the set of American voters. But probabilities in the modal sense are present in the model. They are needed to define the assumption of *random* sampling (one in which all American voters have an equal probability of being included in the sample) or to correct for non-randomness, and to correct for other biases, like the probability that someone would vote Democrat given that they say that they would vote Democrat.

A third reason for the mistake arises because the *evidence* for probabilistic hypotheses always derives from a *population* of actual events. To test a hypothesis, it is not enough to check that the cause co-occurs with the effect. Rather, we must look at the pattern of occurrences in a large number of relevantly similar situations. But it would be a deep conceptual error to conclude that probability is therefore nothing more than the proportions of event types within this set of repeated trials. Multiple trials are actually modeled by defining probabilities over the Cartesian product of the event spaces for each trial.[10] A common *assumption* is that the same probability distribution applies to each trial and all trials are identical and probabilistically independent (the so-called i.i.d. assumption). With this assumption, the αs and the βs are not only constant within the space of possibilities, but also over the population of instances.

The familiarity of the i.i.d. assumption encouraged logical positivists to explore the idea that probabilities might be defined by their connections with the relative frequencies of actual events in large samples, as exemplified by the law of large numbers and other convergence theorems. But these theorems depend on the simplifying assumptions already mentioned. So the positivist programme would only have defined probabilities in a special case, which is why a *general* definition of probability is unavoidably modal in character. For the same reason, the content of any probabilistic hypothesis is strongly modal.

Having established that the probabilistic assumption made about the residue is crucial to the content of a causal model, let me state the assumption more carefully:

Assumption 1: The mean, or expected value, of the residue is zero for every value of the independent variable X.[11] That is to say, the model equation stripped of its residue term is the *mean curve* of Y given X. Formally, the assumption is that, for all x, $E[U/X=x]=0$, or equivalently, $E[Y/X=x]$ = $\alpha_0+\alpha_1 x$. I will use this assumption because it extends most naturally to the nonlinear case.

5 Dichotomous Causal Variables

Philosophers, and computer scientists, tend to be trained in formal logic, and for that reason they standardly study probabilistic causality in the special

[10] See Hitchcock (1993a) for a more thorough discussion of this point.

[11] The expected value of a random variable, or any function of random variables, is the average or mean value weighted according the probability of those values.

case in which all causal variables take on one of two possible values. These are called *dichotomous*, *binary*, or *yes-no* variables. Equations like X = 1 and X = 0 can be described as events A and not-A, respectively. The purpose of this section is to explain how this special case falls within the more general framework I am considering. Therefore, all the lessons of this paper apply to the wealth of philosophical literature on this subject (e.g., Eells 1991).

Consider the special case of Model 1 in which X and Y are dichotomous yes-no variables. That is, assume that the only possible events are X = 0, X = 1, Y = 0, and Y = 1. The values of 0 and 1 are chosen arbitrarily. Nothing essential depends on this choice. In addition, Model 1 assumes the same basic equation as before:

$$Y = \alpha_0 + \alpha_1 X + U \ .$$

However, U now takes on four possible values: $\{1-\alpha_0-\alpha_1,\ -\alpha_0-\alpha_1,\ 1-\alpha_0,\ -\alpha_0\}$, depending on which the four possibilities hold: X = 1 and Y = 1, X = 0 and Y = 1, X = 1 and Y = 0, or X = 0 and Y = 0, respectively. As before, we assume that the expected value of U is zero no matter what which value of X holds. In particular, the expected value of Y given X = 1 is $\alpha_0 + \alpha_1$, and the expected value of Y given X = 0 is α_0. If we now solve for the unknowns α_0 and α_1, we prove that:

$$\alpha_1 = E(Y/X=1) - E(Y/X=0),$$

and

$$\alpha_0 = E(Y/X=0).$$

But if the expected values are related to probabilities by $E(Y) \equiv 1.P(Y{=}1) + 0.P(Y{=}0)$, then:

$$\alpha_1 = P(Y=1/X=1) - P(Y=1/X=0),$$

and

$$\alpha_0 = P(Y=1/X=0).$$

Therefore the constants (or parameters) of the model are expressed in terms of probabilities. Since the constants (often called path coefficients) represent the causal structure imposed by the model, this proves that the causal structure is related to the probabilities in an interesting way.

In standard statistical terminology,

$$r(X \to Y) = P(Y=1/X=1) - P(Y=1/X=0)$$

is called the *regression coefficient* of Y on X. In the general case of non-dichotomous variables, the regression coefficient has a more general expression, but it is still a function of the probabilities and it is still equal to the constant α_1.

The assumption that α_0 and α_1 are constants implies that the forward probabilities, $P(Y=1/X=1)$ and $P(Y=1/X=0)$ are *constants*. The way I describe it, the forward-directed probabilities are physical properties, or propensities, of the system because α_0 and α_1 represent physical properties of the system. It is quite standard in physics and biology that model parameters represent properties such as mass, charge, or fitness.

In causal modeling, the parameters represent the causal structure, which are a function of the forward-directed probabilities. On the other hand, the backward-directed probabilities $P(X=1/Y=1)$ and $P(X=1/Y=0)$ are not constants of the model. Using Bayes Theorem, and the model equation, we can show that,

$$P(X=1/Y=1) = (\alpha_0+\alpha_1)\,P(X=1) \big/ \left[\alpha_0 + \alpha_1 P(X=1)\right]$$

.So the backwards probabilities depend on the constants of the model *together with* the probability $P(X=1)$, but this probability is quite unconstrained by the model (the principle of functional autonomy again). Therefore, according to the model, there is no reason to expect the backwards probabilities to be remain the same from one situation to the next. There is an asymmetry between backwards and forwards probabilities implied by the model.

One such example appears in Sober (1994). The probability that two heterozygote parents (Aa) giving birth to a heterozygote offspring is determined by the laws of genetics to be ½, the probability that a heterozygote offspring had two heterozygote parents is not determined by those same laws.

Another illustration is provided by a modified version of the light switch example. In our house, a pair switches controls a single hall light. The light is on if the switches are up-down or down-up, and off if the pair is up-up or down-down. The causal facts about the wiring fix the probability of the light being on given the switch settings. But no causal facts about *this* system determine the probability that the switch setting is up-down given that the light is on. It depends on the frequency of the switch settings, which has to do with our psychological habits in operating the switches, rather than the wiring itself.

6 Regression Analysis

I will now extend the analysis of yes-no variables to the case of continuous variables. Suppose that the model $Y = \alpha_0 + \alpha_1 X + U$ is the true one, where we make only the weak assumption that the residue has mean zero for all values of *X*. There is a whole set of probability distributions that are compatible with the model. The model does not entail any single one of these probability distributions. It only implies a disjunction of them. If we ignore that fact, and wrongly suppose that the content of the model is captured by a single distribution, then the asymmetry between *X* and *Y* disappears. The reason is that for each distribution there exists an "inverse" mean curve. That is, the probability distribution is also described by an equation $X = \beta_0 + \beta_1 Y + U'$, where U' has mean zero for all values of *Y*. But this symmetry is not a problem for the approach I am developing because the inverse mean curves are different for each probability distribution. That is to say, the β's are not constant, and do not represent physical propensities of the system. The inverse mean curve fails the invariance test.

7 Truly Equivalent Causal Models

Having carefully described the content of causal models, we are in a position to properly define the equivalence of two causal models. As expected, Model 1 and Model 2 will not be equivalent according to this definition. This is a surprising result because, as I will show, my definition of equivalence looks remarkable similar to the definition of VP-equivalence (Definition 1 inVerma and Pearl (1991)), yet Model 1 and Model 2 are not VP-equivalent (as follows easily from their Theorem 1). However, there is a subtle and important difference in the definitions that resolves the inconsistency, as I will point out at the end of the section. I will also argue that their definition does not capture the relevant notion of empirical indistinguishability. To do this I need some rather precise terminological conventions.

A *model equation* is an equation

$$Y = f(\boldsymbol{X}; \mathrm{a}) + U,$$

where $\boldsymbol{X}$ stands for a list of variables, α is a set of parameters, and *f* is any function of those variables and parameters, linear or nonlinear, and *U* is the residue term. $f(\boldsymbol{X}; \mathrm{a})$ may be thought of as a *family* of functions $g(\boldsymbol{X})$, where each member of the family is picked our by a specific numerical assignment of values α. The equation of Model 1,

$$Y = \alpha_0 + \alpha_1 X + U_Y,$$

is an example of a model equation. In this example, $\boldsymbol{X}$ is a single variable *X*, and α is a pair of parameters (α_0, α_1). The specific equations $Y = 1 + X + U$ and $Y = -3 + 2X + U$ are different members of the family. A *causal model* is a model equation together with the assumptions made about the residue.[12] I will adopt Assumption 1 in section 4 as the assumption to make about the residue. I will also assume that there is a single, fixed, distribution for the residue that applies to all instances of the model equation. None of these assumptions are cast in stone, but they are fairly typical.

Now consider a particular *member* of the family $Y = f(\boldsymbol{X}; \mathrm{a}) + U$. That is, consider $Y = f(\boldsymbol{X}; \boldsymbol{v}) + U$, where $\boldsymbol{v}$ is a particular number assignment of values to the parameters α. Let us refer to this equation, combined with the assumption about the residues, as a *causal hypothesis*, as opposed to a causal model (so a causal model is a family of causal hypotheses).

[12] This definition is easily extended to the case in which there is more than one model equation, for more that one dependent variable, as there would be for the model in Figure 1, for example. But, since the issues of this paper are amply explained in terms of simple causal models, I will postpone the general case for another time.

The content of a causal hypothesis may be represented as a set of probability distributions over the possible values of the variables Y and $\boldsymbol{X}$ that meet the assumptions of the hypothesis. That is, a causal hypothesis may be represented by a set of probability distributions $p(\boldsymbol{x},y)$, where $\boldsymbol{x}$ and y range over the possible values of $\boldsymbol{X}$ and Y, respectively. Denote the causal hypothesis by H. Then under the assumptions we have made about the residue term, there will exist a unique conditional probability function $p_H(y/\boldsymbol{x})$, such that such that $p(\boldsymbol{x},y)\in H$ if and only if

$$p(\boldsymbol{x},y)/p(\boldsymbol{x}) \ = \ p_H(y/\boldsymbol{x}),$$

for all $\boldsymbol{x}$ and y. In words, $p(\boldsymbol{x},y)$ is in the family of probability distributions representing H if and only if the conditional probabilities it defines match the ones *predicted* by the hypothesis. This constraint does not determine a unique probability distribution, so every causal hypothesis is represented by a non-singleton family of probability distributions. Logically speaking, H may be thought of as asserting that, in any given situation, the true probability distribution is one of those in H. H is a huge disjunction in other words. A probabilistic property is predicted, as opposed to accommodated, by H if and only it is *invariant* feature of all probability distributions in H.

For example, Model 1 predicts the forward conditional probabilities. Model 1 can *accommodate* backwards probabilities, but they does not predict them. The opposite is true for Model 2. That, for us, is a very important distinction.

Our formal representation of causal models is complicated by the fact that a model is a set of hypotheses and a hypothesis is a set of probability distributions. It follows that a causal model is a set of sets of probability distributions. Verma and Pearl (1991) appear to dump all the probability distributions into one superset, but if you do that, then the distinction between Model 1 and Model 2 will disappear. I suspect that this is where Verma and Pearl (1991) go wrong.

Two causal hypotheses are *equivalent* if and only they are represented by exactly the same set of probability distributions. Two causal models, M_1 and M_2, are *equivalent* if and only each hypothesis in M_1 is equivalent to some hypothesis in M_2 and each hypothesis in M_2 is equivalent to some hypothesis in M_1. That is the correct definition of equivalence.

On this definition, Model 1 and Model 2 are not equivalent. In fact, they a disjoint, for there is no hypothesis in Model 1 that is equivalent to any hypothesis in Model 2, because Model 1 does not predict backward probabilities and Model 2 does not predict forward probabilities.

Given that this result conflicts with a large body of literature on automated causal inference, let me spell it out in the simplest case imaginable. In the smoking example, let S = person x is a smoker, $\overline{S}$ = person x non-smoker, D = person x develops heart disease, and $\overline{D}$ = person x does not develop heart disease. Let M_1 be the model that says that smoking cause heart disease, and M_2 the model that says heart disease causes (prior) smoking. Let $H_1(\theta, \lambda)$ be a causal hypothesis in M_1, where θ and λ are constants that define the forward probabilities. That is, $P(D/S)=\theta$ and $P(D/\overline{S})=\lambda$. Now, each probability distribution in $H_1(\theta, \lambda)$ will assign numbers to the probabilities of S & D, S & $\overline{D}$, $\overline{S}$ & L, and $\overline{S}$ & $\overline{D}$ in the following way:

$$P(S\&D) \ = \ \theta P(S),$$

$$P(S\&\overline{D}) \ = \ (1-\theta)P(S),$$

$$P(\overline{S}\&D) \ = \ \lambda(1-P(S)) \text{ and}$$

$$P(\overline{S}\&\overline{D}) \ = \ (1-\lambda)(1-P(S)).$$

There are many such probability distributions in $H_1(\theta, \lambda)$ because there are many ways of assigning numerical values to $P(S)$. However, all the probability distributions in $H_1(\theta, \lambda)$ share the property that $P(D/S)=\theta$ and $P(D/\overline{S})=\lambda$. These are the probabilities that $H_1(\theta, \lambda)$ predicts. M_1 is not equivalent to M_2 because $H_1(\theta,\lambda)$ is not equivalent to any hypothesis in M_2.

Verma and Pearl (1991, p. 256), from what I can gather from the paper, define equivalence in the following way (translated into my terminology). M_1 is *VP-equivalent* to M_2 if and only if every probability distribution in some hypothesis in M_1 appears in some hypothesis in M_2 and *vice versa*. On this definition, M_1 is VP-equivalent to M_2 because any probability distribution determined by

$$P(D/S) = \theta,\ P(D/\overline{S}) = \lambda \text{, and } P(S) = \rho,$$

will be the same as the probability distribution determined by

$$P(S/D) = \theta\rho / (\theta\rho + \lambda(1-\rho)),$$

$$P(S/\overline{D}) = (1-\theta)\rho / ((1-\theta)\rho + (1-\lambda)(1-\rho)),$$

$$P(D) = \theta\rho + \lambda(1-\rho),$$

and *vice versa*. In words, M_1 is VP-equivalent to M_2 because a probability distribution is *accommodated* by M_1 if and only if it is *accommodated* by M_2. They have failed to respect the distinction between prediction and accommodation.

The received approach to causal inference appears to view it as an inductive procedure that takes us from empirical data to a causal *model*. However, it is better to view causal inference as an inductive inference from data to causal *hypothesis*. If we are interested in causal models, then the causal model can be inferred deductively from that point. So, if we focus on the equivalence of inductive *hypotheses*, rather than models, we are not going to get muddled about sets of sets probability distributions. It seems that everyone should agree that two causal hypotheses are equivalent if and only if they share the same probability distributions. In fact, the formal definition of equivalence in Verma and Pearl (1991) looks just like my definition of equivalent causal hypotheses. But somewhere along the line, they fail to keep the distinction between hypotheses and models clearly in mind (they use these terms differently, but our differences are not merely terminological).

However, there is a sophisticated defense of their approach, which argues that the notion of accommodation is actually the right one. They might argue that the distinction between models and hypotheses not relevant. Here is an argument: Causal inference is a two-step process. The first step is to infer a single probability distribution that best represents the empirical frequencies in the data. All causal conclusions must be based on this inferred probability distribution, which takes us to the second step of causal inference. There is nothing else on which to base the inference. Call the probability distribution Q. The best fitting hypothesis in M_1 will succeed in predicting some features of Q and in accommodating others. But likewise, the best fitting hypothesis in M_2 will succeed in predicting just as many features of Q and in accommodating the rest. To assume that some features of Q are more worthy than others for prediction as opposed to accommodation is the beg the question in favor of one model and against the other.

The remainder of this paper is devoted to showing that this argument is mistaken. I think that their picture view of causal inference is fundamentally wrong.

8 The Consilience of Causes

In section 4, I explained the sense in which the contents of Model 1 and Model 2 are different. I explained the difference in modal terms—in terms of the content of single-case probabilities. However, no models are distinguishable by any observations made in a single case. That is a universal fact, true of any kind of model in science. One has to examine numerous instances that are assumed to be identical in relevant respects. So, it is necessary to look at a population of n events

$$\{(X_1 = x_1, Y_1 = y_1), (X_2 = x_2, Y_2 = y_2), \ldots, (X_n = x_n, Y_n = y_n)\}$$

Now, we need to further assume that the same model applies to each instance. That is, we must compare Model 1 applied to n pairs of events, against Model 2 applied to n pairs of events. But this is still not sufficient, and to understand why it is not is to understand why one must respect the distinction between a causal model and a causal hypothesis is vitally important in causal inference. The reason is that the weak assumption that the same model applies in each instance allows that different hypotheses in the model apply in each instance. Or to put it another way, it allows that different values of the parameters apply in each instance. This means that there are at least as many adjustable parameters as there are data points, and we are no better off than we were in the single instance. There is no test of the model at all. Therefore, when comparing two models, we must assume that the same causal *hypothesis* applies in each instance. That is, we must assume that the parameters are *constants*. Then, and only then, do we have a crucial test of the models.

This bring us back to the assumption of identical independent distributions (i.i.d.) referred to earlier. The ‘identity’ part of the assumption implies not only that the same model applies to each instance, but also that the same causal hypothesis applies in each instance. *Without this assumption*, or something like it, *there would be as many parameters as there are data, and it would be impossible to estimate their values accurately*.

However, it appears that Verma and Pearl (1991) may circumvent this difficulty by first fitting a single probability distribution to the data. For in that estimation problem, it is tacitly assumed that a single probability distribution and therefore a *single* set of parameter values apply to all the data. That is, i.i.d. assumption is built into this first step. It follows that, although the role of the i.i.d. assumption is crucial to my point, it does not by itself prove that Model 1 and Model 2 are *empirically* distinguishable.

So, let me push the question one step further. Suppose one scientist does a regression of Y on X and find a value of the parameters α_0 and α_1 that best fit the data, while a second scientist does a regression of X on Y, and finds values for β_0 and β_1 that best fit the data. The degree of fit, in a sense, is the test of the model. But is it a crucial test? That is, could this procedure provide evidence for Model 1 that counts against Model 2. The surprising answer is ‘no’.

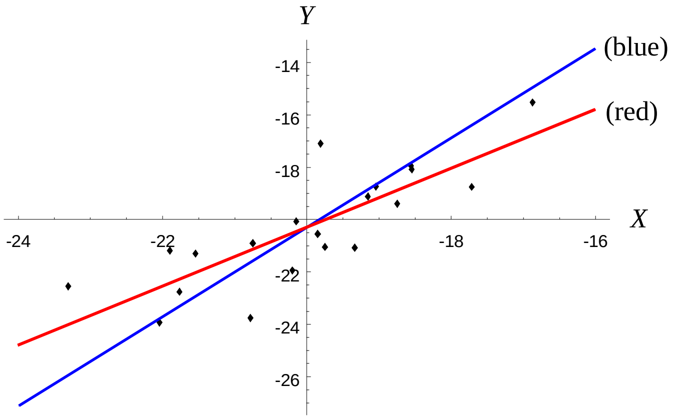

Figure 6: The line with the lesser slope (red) is the standard regression curve of *Y* on *X*, while the other (blue) is the regression curve of *X* on *Y*.

In fact, going by the degree of fit lead you to the wrong hypothesis, as the following simulations show. I generated 20 data points from a probability distribution given by $Y = X + U$, where U is a normally distributed residue with mean zero and variance 2, and X is normally distributed with mean -20 and variance 4. The data points are shown in Figure 6. Then I did a least squares linear regression of Y on X, which is the (red) line in Figure 6 with the lesser slope. Finally, using the same data, I plotted a ‘backwards’ least squares regression taking X to be the dependent variable and Y as the independent variable. The resulting best fitting curve is the (blue) line with the greater slope plotted in Figure 6. The surprising fact is that the backwards regression fitted better than the forwards regression, even though the forwards model is the true on in this case. I verified that this was not an accidental feature of this par-

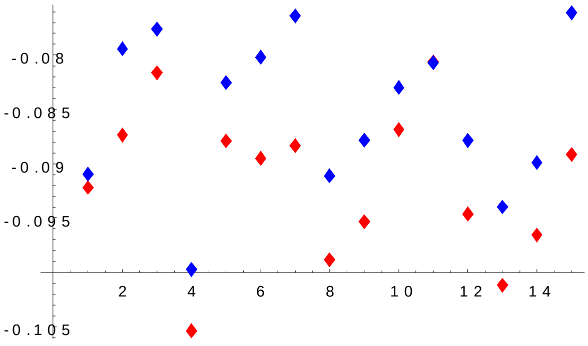

Figure 7: The degrees of fit for the best fitting forwards regressions (red) and backwards regressions (blue) for 15 different data sets of 20 points each. The backward regression fitted best in each case.

ticular data set by plotting a further 15 data sets like this one randomly generated from the same distribution. The backwards regression consistently fitted better than the forwards regression in every case. The fits are plotted in Figure 7.

While the backwards regression curves fit the seen data better, they are less predictively accurate. This shown in Figure 8, where I have plotted the regression curves for all 15 trial mentioned above. The true curve passes through the origin, and the predictive accuracy of a regression curve is measured by it closeness to the true curve (see Forster and Sober (1994) or Forster (forthcoming) for a more detailed discussion of predictive accuracy). The forward regression curves pass fairly close to the origin, but the backward regression curves, with one exception, miss their mark by a wide margin.

The example shows that a naïve empirical criterion of model selection that says “choice the model that fits case” will not work well. The traditional response to this fact is to amend the naïve criterion to include simplicity as a factor in model selection. The problem is to say what simplicity is, and how it is factored into a model selection criterion. There is a huge body of literature on this problem (see Forster and Sober (1994), or Forster (forthcoming) for an introduction). The basic theoretical idea is that the best fitting hypothesis in a model is subject to two kinds of error. The first is called *model bias*. The accuracy of a best fitting case is clearly limited by what is available in the model to work with. A best fitting curve cannot be close to the true curve if there are no curves in the model close to the true curve. In such a case, we say that the model is misspecified or biased (Kruse (1997) uses the term ‘model error’). In our example, Model 1 is unbiased because it contains the true hypothesis. Model 2 is not very biased either because it contains a curve that passes through the origin (it may not get the probability distribution for the residue exactly right, but my example here is designed so that it could do a good job here as well, as we will see in a moment).

A second source of error is the estimation error (Kruse (1997)), the overfitting error, or the more simply, the variance of the estimates. This is a kind of sampling error in which random fluctuations experienced in small data sets will lead random variation in the estimated values of the model parameters. All of the well known model selection criteria assume that this kind of error increases proportionally to the number of adjustable parameters in a model (or the dimension of the model—see Forster (submitted)). The number of adjustable parameters is therefore the relevant measure of simplicity. In the present example, both models have the same number of adjustable parameters (two). And, indeed, from Figure 8, we see that both models display roughly the same variation around a central mean value for their parameter estimates. However, the simplicity term will have no effect on the comparison of Model 1 and Model 2 in this example because the simplicity terms will cancel out in all of the standard model selection.

The problem is that, in the case of the backwards regression, there is a third source of error, which is not accounted for in any of the standard model selection methods. There is a *systematic* estimation error whereby the mean parameter values are not centered on the values that are the best in the model. The problem of *causal* model selection, therefore, encounters a new kind of model selection

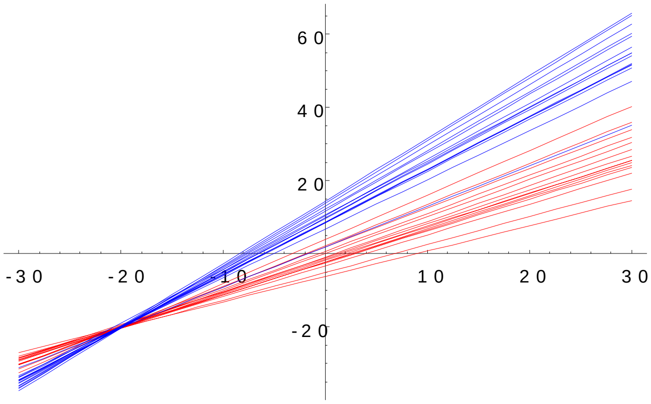

Figure 8: While the true curve passes through the origin, the fitted regression curves miss the mark. However, the backwards regression curves (blue) are systematically innacurate in a way in which the forward regression curves (red) are not.

problem.

What is the solution to this problem? One idea is that we need to use varied data (see Kruse (1997) for an excellent discussion of how this affects the maximization of predictive accuracy). For example, suppose we supplement the data shown in Figure 6, centered at the point (–20, –20), with data centered at the point (+20, +20). The new more varied data set is shown in Figure 10. In this case, both the forwards and backwards regression lines will fit closely to the true curve—I have verified that they are very close in fit and in predictive accuracy. However, this does not solve our problem because we have not arrived at a way of favoring the true model. All we have done is to removed a bias that favors the wrong model.

Before offering a solution to this problem, let me say why it is important. One might argue that the use of varied data shows that our problem is unimportant. So what if we can't use varied data to discriminate in favor of the true model? If closeness to the truth, or predictive accuracy, is the principal goal to which scientists aspire (Forster (forthcoming), Forster and Fitelson (unpublished), Sober (unpublished)), then I have shown that either model will serve equally well if we use varied data. Nothing else matters.

The reply to this objection, note that predictive accuracy is always relative to a domain (Forster (forthcoming)). In our example, the data centered at the point (–20, –20) may be thought of as belonging to one domain, while the data centered at (+20, +20) belongs to a second domain. To take predictive accuracy as a goal is to aim to maximize predictive accuracy in all domains. After all, it was the failure of regression curves fitted in Domain 1 (Figure 9) to fit the data in Domain 2 that motivates the move to varied data in the first place. So the goal of predictive accuracy is an open-ended kind of goal (like truth) because it is meant to apply to an open-ended list of domains. The difference between Model 1 and Model 2 will show up in other domains, even if it does not show up in the combined domains of Domain 1 and Domain 2. In causal models, we want to predict what will happen under conditions of manipulation of various kinds. If we manipulate the variable *X*, then Model 1, being the true model, will correctly predict the response of *Y*. But Model 2 does not predict that *Y* will respond in the same way, because it contends that *X* does not cause *Y*. This difference is a difference in predictive accuracy.

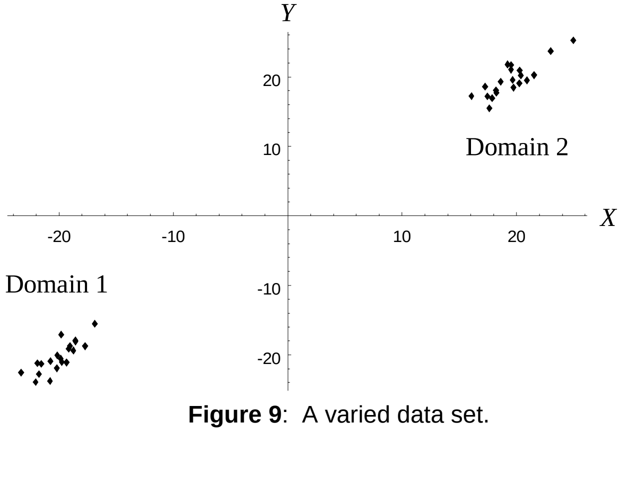

Figure 9: A varied data set.

Of course, we could carry out such manipulations (if we are able), and discriminate between the models on that basis. But our question is whether a *prior* discrimination is possible. If there is, then we want to know about it.

An affirmative answer is already apparent in Figure 8. For if we *test* the regression curves fitted to the data in Domain 1 against the data in Domain 2, we see that Model 1 does significantly better than Model 2. The fitted curve in Model 1 will pass quite close to the data in Domain 2, while Model 2 will systematically miss its mark, usually by quite a wide margin. The solution is very familiar to philosophers of science—we should test the models. This kind of test is known as *cross validation* in the model selection literature, although it is a little different from the standard version of this criterion.[13] The idea behind cross validation is that we should divide the seen

[13] The method was developed by Mosteller and Wallace (1963) to determine the authorship of disputed *Federalist* papers. An important theorem was proven by Stone(1977), who showed that leave-one-out cross validation is asymptotically equivalent to Akaike's AIC method. Also see Turney (1994).

data into two subsets. In machine learning, the first subset is called the *training data*, and the second is called the *test data*. However, there are two standard refinements of the idea. The first arises from the fact that, once the a model is selected, we should make full use of the seen data in fitting the model, and this is clearly correct. It seems that we should make the training set as close as possible to the full set, so that the tested curve is as close as possible to the curve we will eventually use for prediction. So, the training data is taken to be the full data *with one data point left out*. But then we are only testing against one data point, and our test will be unduly subject to random sampling error. It appears that the solution is to repeat the leave-one-out test *n* times for each data point, and averaging the results. The model with the highest average score wins. However, this standard form of cross validation will not solve our problem in our example because the tested curves will very close to the true curve in every case. We need a non-standard form of cross validation.

In our application, the data in Domain 1 is of a *different kind* from the data in Domain 2 because they cover different ranges of values for the variables (both *X* and *Y*, so the division into kinds does not beg the question against either of the competing models). We should fit the models in Domain 1 and test against Domain 2, and then fit the models to Domain 2 and test against the data in Domain 1. The test appears to be the most discriminating because it tests the ability of the model to 'reach out' from one domain into another. It tests for extrapolation rather than interpolation. In our setup, this test favors Model 1 over Model 2 (Figure 8). Symmetry considerations show that the opposite would be true if the data were generated by Model 2 rather than Model 1.

There is a second way of performing the same test. First, fit each model separately to the two domains of data. We will end up with two sets of parameter estimates for each model. Then score the models according to whether their parameters estimated in one domain agree with the estimates obtained from the other domain. The results will be the same as before.

In the philosophy of science literature this kind of test is referred to as the "agreement of independent measurements." For example, Newton was able to infer the mass of the Earth from terrestrial motion and then estimate the same quantity from the motion of the moon, and other phenomena, such as the patterns of the tides. The agreement of these independent measurements was impressive evidence for Newton's theory of gravitation, for it provided direct evidence that terrestrial and celestial phenomena resulted from a single cause. Newton captured this intuition in his famous first rule of reasoning (Book III of his *Principia*): "We are to admit no more causes of natural things than such as are both true and sufficient to explain their appearances." In reference to this rule, an influential nineteenth century historian of science claims that:

> When the explanation of two kinds of phenomena, distinct, and not apparently connected, leads us to the same cause, such a coincidence does give a reality to the cause, which it has not while it merely accounts for those appearances which suggested the supposition. This coincidence of propositions inferred from separate classes of facts, is exactly what we noticed in the *Novum Organon Renovatum*, as one of the most decisive characteristics of a true theory, under the name of *Consilience of Inductions*.
>
> That Newton's First Rule of Philosophizing, so understood, authorizes the inferences which he himself made, is really the ground on which they are so firmly believed by philosophers. (William Whewell, in Butts (1989), p.330)

Notice that when Newton and Whewell talk about explaining phenomena in terms of a common cause, they are not talking about explaining two dependent variables in terms of a single independent variable. Their talk of 'cause' might be better translated as 'law'. In Newton's case, the very same law, the inverse square law of gravitation, with exactly the same parameter value (the gravitational mass of the Earth) successfully explains two different kinds of phenomena—terrestrial phenomena and celestial phenomena. This is exactly analogous to the cross validation test I have been describing, which is why I will refer to the test as the consilience of inductions from this point forward.

Finally, I note that this kind of test is not new to causal modeling. As Wright (1923, p.241) puts it, “if the logical consequences [of a causal hypothesis] can be shown to agree with independently obtained results, it contributes to the demonstration of the truth of the hypothesis in the only sense which can be ascribed to the truth of a natural law.”

9 Discussion

In summary: The consilience of inductions (note the plural) requires (1) that we divide the total data set into separate subsets, where each subset is of a different kind. Then (2) we perform an induction on each of the subsets separately. By ‘induction’, Whewell is referring to the process of fitting the model to the data. Finally, (3) we score a model according to how well the two parts “jump together” (as Whewell put it). This may be measured by looking at either the prediction error on the data not used in the construction of the hypothesis or the extent to which the parameter values agree.

How does the consilience of inductions tie in with our previous distinction between prediction and accommodation? Well, the consilience of inductions is a direct test of ability of a model to *predict* phenomena in one domain from data in a different domain. Consider a comparison between Newton’s gravitational model (N) against a conjunction of Galileo’s law applied to terrestrial motion, and Kepler’s laws applied to the motion of the moon (G & K). N and G & K are able to accommodate the total data (to the level of approximation appropriate in this example). However, the disunified hypothesis G & K is unable to predict the moons motion by fitting to terrestrial motion, or vice versa. It has the power of accommodation, but not of prediction. That is to say, there is no consilience of inductions in G & K to match the consilience of inductions for N.

The simulated example in the previous section is entirely analogous, except for two things. G & K failed the prediction test because it makes no predictions at all. The estimation of parameters in G provides no information about the parameters in K. In contrast, Model 2 does make predictions in Domain 2, but they are not *successful* predictions. The other difference is that G & K is less unified than N (it has more adjustable parameters), whereas Model 1 and Model 2 are equally unified. The two differences are not unrelated.

The simulated example in this section clearly shows that the consilience of inductions can discriminate between two models even when no direct test on the *pooled* data can. This is a surprising result to many philosophy of science, who have subscribed to the *principle of total evidence*, which says that we should make use of all known evidence when deciding between theories. For the consilience of inductions appears to violate this principle by using only part of the data to construct the hypothesis and part of the data to test the hypothesis. Further consideration shows that there is no real violation of the principle because all the data is being used in the test in one way or the other. However, the principle may have had a psychological effect in dissuading us from examining these kinds of tests.

I began this paper with the following picture of causal inference: First, represent the data in terms of a single probability function. Call the probability distribution Q. Second, exploit the conditional independencies in this probability distribution in order to draw the strongest possible causal conclusions. Towards the end of section 7, I considered an argument to a conclusion that appeared to deny that any test like the consilience of inductions could work. The argument went like this: The best fitting hypothesis in Model 1 will succeed in predicting some features of Q and in accommodating others. But likewise, the best fitting hypothesis in Model 2 will succeed in predicting just as many features of Q and in accommodating the rest. To assume that some features of Q are more worthy than others for prediction as opposed to accommodation is the beg the question in favor of one model and against the other. In the simulated example, this conclusion does follow from these premises, for if we obtain a single Q from the *pooled* data, then there was no test that favored the true model over the false one. Or to put it another way, there is no test that favors the true model without begging the question against false one.

But that is not how the consilience of inductions proceeds. If it is to fit into this kind of picture at all,[14] then it would go like this: (1) Represent the data in terms of a two probability functions, Q_1 and Q_2; one for each of the two different kinds of data. (2) Make separate causal inferences from Q_1 and Q_2. (3) Test each of the results to see how well the hypothesis performs on the other domain of data. (4) Infer a composite causal model on the basis of those results.

For each Q_1 and Q_2, there is a distinction being made between aspects that it is good to predict and those that should be accommodated. Roughly speaking, it is good to predict similarities, and good to accommodate the differences. But this distinction does not beg the question for or against any particular hypothesis because the differences and similarities between Q_1 and Q_2 are determined by the data.

There are two more important advantages of the method as well:

- It respects the principle of functional autonomy (section 3), which states that relationship of a variable on set of causes is independent of the relationship amongst those causes. For even if the relationship amongst the causal variables are clearly different in Q_1 and Q_2, the consilience of inductions can still test for the similarities predicted by a model.
- The test is less sensitive to the existence of 'latent' common causes. For if the 'latent' common causes affect only the relationship amongst the causes, then the dependence of the effect variable on those causes is unaffected.

The automated reasoners from the computer sciences may grumble that this talk of "different kinds of data" is vague and unhelpful. It is vague, but it has not been unhelpful to scientists. It is unhelpful to automated reasoners if they do not know how to program a computer to divide the total data appropriately. I am not claiming, by any means, that this cannot be done. My conclusion is merely that causal inference is harder than is standardly thought. And, as for many hard kinds of inference, subject to a complex array of errors, scientists have adopted methodological principles that manage these errors (even if they have not analyzed them as such). The example of causal inference promises to be no exception.

Acknowledgments

I am grateful to Marty Barrett, Ellery Eells, Branden Fitelson, Leslie Graves, Dan Hausman, Christopher Hitchcock, Gürol Irzik, Joonsung Kim, Elliott Sober, Chris Stephens, and Jim Woodward for helpful discussions on related issues. I would also like to thank Dan Hausman for his encouragement, without which this paper would not have been written at this time.

Department of Philosophy
5185 Helen C. White Hall
University of Wisconsin
Madison, WI 53706
USA

[14] I don't happen to believe that this picture is accurate, because the inferred Q will depend, in part, on the model under consideration. Nevertheless, it is a useful idealization for the purpose of understanding the logic of confirmation.

References

Arntzenius, Frank (1993): "The Common Cause Principle." *PSA 1992 Volume* **2***:* 227-237. East Lansing, Michigan: Philosophy of Science Association.

Arntzenius, Frank (1997): "Transition Chances and Causation." *Pacific Philosophical Quarterly* **78**: 149-168.

Bogen, James & James Woodward (1988): "Saving the Phenomena." *The Philosophical Review,* vol. **XCVII:** 303-352.

Butts, Robert E. (ed.) (1989). *William Whewell: Theory of Scientific Method.* Hackett Publishing Company, Indianapolis/Cambridge.

Cartwright, Nancy (1983): *How the Laws of Physics Lie.* Oxford: Oxford University Press.

Eells, Ellery (1991): *Probabilistic Causality.* Cambridge: Cambridge University Press.

Forster, Malcolm R. (1984): *Probabilistic Causality and the Foundations of Modern Science.* Ph.D. Thesis, University of Western Ontario.

Forster, Malcolm R. (1988): "Unification, Explanation, and the Composition of Causes in Newtonian Mechanics." *Studies in the History and Philosophy of Science* **19:** 55 - 101.

Forster, Malcolm R. (forthcoming): "The New Science of Simplicity" *Proceedings of the Tilburg Conference on Simplicity, 9-11 January 1997,* edited by Hugo Keuzenkamp, Michael McAleer, and Arnold Zellner.

Forster, Malcolm R. (submitted): "The Gruesome Curve-Fitting Problem." *The British Journal for the Philosophy of Science*.

Forster, Malcolm R. and Branden Fitelson (unpublished): "Do Scientists Value Theories for their Truth or for their Predictive Accuracy? A Bayesian Analysis."

Forster, Malcolm R. and Elliott Sober (1994): "How to Tell when Simpler, More Unified, or Less *Ad Hoc* Theories will Provide More Accurate Predictions." *The British Journal for the Philosophy of Science* **45:** 1 - 35.

Glymour, Clark (unpublished): "Learning Causes: Psychological Explanations of Causal Explanation."

Harper, William L. (1989): "Consilience and Natural Kind Reasoning." In J. R. Brown and J. Mittelstrass (eds.) *An Intimate Relation:* 115-152. Dordrecht: Kluwer Academic Publishers.

Hausman, Daniel M. (1996): *Causal Asymmetries*. Unpublished manuscript.

Hitchcock, Christopher R. (1995): "The Mishap at Reichenbach Fall: Singular vs. General Causation," *Philosophical Studies* **78:** 257-291.

Hooker, Cliff A. (1987): *A Realistic Theory of Science*. Albany: State University of New York Press.

Hoover, Kevin (1993): "Causality and Temporal Order in Macroeconomics or Why Even Economists Don't Know How to Get Causes from Probabilities." *British Journal for the Philosophy of Science* **44**: 693-710.

Hoover, Kevin (1994): "Econometrics as Observation: the Lucas Critique and the Nature of Econometric Inference." *Journal of Economic Methodology* **1**: 65-80.

Humphreys, Paul and David Freedman (1997): "The Grand Leap." *British Journal for the Philosophy of Science* **47**: 113-123.

Irzik, Gürol (1996): "Can Causes be Reduced to Correlations?", *British Journal for the Philosophy of Science* **47:** 249 - 270.

Korb, Kevin B. and Chris S. Wallace (1997): "In Search of the Philosopher's Stone: Remarks on Humphreys and Freedman's Critique of Causal Discovery," *British Journal for the Philosophy of Science* **48**: 543-553.

Kruse, Michael (1997): "Variation and the Accuracy of Predictions." *British Journal for the Philosophy of Science* **48**: 181-193.

Mosteller, Frederick, and David L. Wallace (1963): "Inference in the Authorship Problem: A comparative study of discrimination methods applied to the authorship of the disputed *Federalist* papers." *Journal of the American Statistical Association*. **58**: 275-309.

Papineau, David (1989): “Pure, Mixed and Spurious Probabilities and Their Significance for a Reductionist Theory of Causation,” in Kitcher and Salmon (1989). pp. 307 - 348.

Papineau, David (1991): “Correlations and Causes.” *British Journal for the Philosophy of Science* **42**: 397- 412.

Papineau, David (1993): “Can We Reduce Causal Direction to Probabilities?” In D. Hull, M. Forbes, and K. Okruhlik (eds.) *PSA*, Vol. 2. Eest Lansing, Philosophy of Science Association, pp. 238-252.

Redhead, Michael (1989): “Nonfactorizability, Stochastic Causality, and Passion-at-a-Distance”, in J. Cushing and E. McMullin (eds.) *Philosophical Consequences of Quantum Theory*. Notre Dame: University of Notre Dame Press.

Robins, James M. and Wasserman, Larry (1998): “On the Impossibility of Inferring Causation from Association without Background Knowledge.” In *Computation and Causality*, eds. Glymour, C., and Cooper, G., AAAI/MIT Press, to appear.

Salmon, Wesley (1971): *Statistical Explanation and Statistical Relevance.* University of Pittsburgh Press.

Skyrms, Brian (1980): *Causal Necessity: A Pragmatic Investigation of the Necessity of Laws*. Yale University Press, New Haven.

Simon, Herbert (1953): “Causal Ordering and Identifiability,” reprinted in *Models of Man*. Wiley, pp. 10-35.

Sober, Elliott (1994): “Temporally Oriented Laws,” in E. Sober (1994) *From A Biological Point of View—Essays in evolutionary philosophy*, Cambridge University Press, pp. 233 - 251.

Sober, Elliott (unpublished): “Instrumentalism Revisited.”

Spirtes, Peter, Clark Glymour and R. Scheines (1993*): Causation, Prediction and Search*. New York: Springer-Verlag.

Spirtes, Peter, Clark Glymour and R. Scheines (1997*)*: “Reply to Humphreys and Freedman’s Review of *Causation, Prediction, and Search*” *British Journal for the Philosophy of Science* **48**: 555-568.

Stone, M. (1977): An Asymptotic Equivalence of Choice of Model by Cross-validation and Akaike’s Criterion.” *Journal of the Royal Statistical Society B* **39**: 44-47.

Turney, Peter D. (1994): “A Theory of Cross-Validation Error.” *The Journal of Theoretical and Experimental Artificial Intelligence* **6:** 361-392.

Verma, T. S. and Judea Pearl (1991): “Equivalence and Synthesis of Causal Models,” in P. P. Bonissone, M. Henrion, L.N. Kanaland J.F. Lemmer (eds.) *Uncertainty in Artificial Intelligence* **6**. Amsterdam, Elslevier Science Publishers, pp. 255-268.

Woodward, James (1993): “Capacities and Invariance,” in J. Earman, A. Janis, G. Massey, and N. Rescher, eds. *Philosophical Problems of the Internal and External Worlds: Essays Concerning the Philosophy of Adolph Grünbaum*. Pittsburgh: University of Pittsburgh Press, pp. 283-328.

Woodward, James (1997): “Explanation, Invariance and Intervention.” *Philosophy of Science* **64**.

Woodward, James (1998): “Causal Models, Probabilities and Invariance.” Forthcoming *Proceedings of the Notre Dame Conference on “Causality in Crisis*.”

Wright, Sewell (1921). “Correlation and Causation,” *Journal of Agricultural Research* **20:** *557-585*.

Wright, Sewell (1923). “The Theory of Path Coefficients: A Reply to Niles’s Criticism,” *Genetics* **8:** 239-255.