

Causal reasoning and backtracking

James M. Joyce

Published online: 15 October 2009
© Springer Science+Business Media B.V. 2009

Abstract I argue that one central aspect of the epistemology of causation, the use of causes as evidence for their effects, is largely independent of the metaphysics of causation. In particular, I use the formalism of Bayesian causal graphs to factor the incremental evidential impact of a cause for its effect into a direct cause-to-effect component and a backtracking component. While the “backtracking” evidence that causes provide about earlier events often obscures things, once we restrict attention to the cause-to-effect component it is true to say promoting (inhibiting) causes raise (lower) the probabilities of their effects. This factoring assumes the same form whether causation is given an interventionist, counterfactual or probabilistic interpretation. Whether we think about causation in terms of interventions and causal graphs, counterfactuals and imaging functions, or probability raising against the background of causally homogenous partitions, if we describe the essential features of a situation correctly then the incremental evidence that a cause provides for its effect in virtue of being its cause will be the same.

Keywords Causal inference · Incremental evidence · Bayesian causal graph · Markov condition · Imaging · Backtracking counterfactual · Probabilistic causation

My topic is the *epistemology* of causation. Philosophical discussions of causation usually focus on metaphysics: Is causation merely constant conjunction? Is it a projection of human habits of inference? Do causal claims presuppose the truth of general laws? What is the relation between causation and our ability to manipulate events? How are causal and counterfactual claims related? While these are important questions, it is no part of my purpose here to answer them. First, I don't

J. M. Joyce (✉)
Department of Philosophy, University of Michigan, Ann Arbor, MI, USA
e-mail: jjoyce@umich.edu

have the answers. Second, one of my themes will be that we can learn a great deal about the *epistemology* of causal claims without committing to any specific metaphysics of causation.

One part of the epistemology of causation involves the *discovery* of causes. Here one infers causal connections from statistical correlations or results of controlled experiments. This will not be discussed here. While I am concerned with beliefs of the form “*c* causes *e*,” for *c* and *e* particular events, I am interested in their role as *premises* of arguments, rather than conclusions. One way in which such causal claims can figure as premises involves inferring the existence of causes from the presence of effects.

- *Effect-to-Cause Inference*: Given that *c* is believed to be a promoting (inhibiting) cause of *e*, under what conditions will learning *e* provide evidence for (against) *c*?

This is not my topic either. Rather, I am interested in

- *Cause-to-Effect Inference*: Given that *c* is believed to be a promoting (inhibiting) cause of *e*, under what conditions will learning *c* provide evidence for (against) *e*.

We often cite causes as evidence for their effects, and I want to understand how this aspect of the epistemology of causation works. I shall argue that the basic logic of cause-to-effect inference is the same on all the main competing metaphysical theories of causation, including probabilistic, counterfactual, and interventionist theories.

My approach to the problem will be the broadly Bayesian one I learned by reading Brian Skyrms’ marvelous books *Causal Necessity* (1980) and *Pragmatics and Empiricism* (1984). Skyrms, building on earlier work of Suppes (1970), Good (1961, 1962), and others, championed a broadly probabilistic understanding of causation (as well as objective chance and counterfactuals). While his view had both metaphysical and epistemological ambitions, I will focus here on only one aspect of Skyrms’ epistemology of causation: namely,

Skyrms’ Thesis: A rational person’s assessment of the degree of causal influence that *c* has over *e* is reflected in her subjective probabilities by the number (in my notation)

$$P^c(e) = \sum_k P(k) \cdot P(e/c \& k)$$

where *P* is the person’s subjective probability, and *k* ranges over a partition *K* of hypotheses whose elements provide a maximally specific description of those features of the world that (i) do not depend causally on whether or not *c* occurs, and (b) are causally relevant to *e*.

As will become clear, I think this approach to the epistemology of causation is importantly right, and that its correctness does not depend on any particular metaphysics of causation.

When one begins to think about the epistemology of causation in probabilistic terms, a natural first suggestion is this:

A (too) Bold Conjecture: Causes *always* provide evidence about their effects. A person who is fully convinced that *c* is a promoting (inhibiting) cause of *e*, and who learns that *c* occurs thereby acquires evidence for believing that *e* will (will not) also occur.

On one Bayesian conception of evidence,¹ which makes confirmation a matter of raising subjective probabilities, this (too) bold conjecture can be rewritten as follows:

Bayesian Version: If a person is convinced that *c* is a promoting (inhibiting) cause of *e*, then her subjective probability for *e* conditional on *c* exceeds (is exceeded by) her unconditional subjective probability for *e*.

This is a subjectivist version of a suggestion, endorsed at various times by Reichenbach (1956), Suppes (1970), and Good (1962), according to which a necessary (but insufficient) condition for *c* to be a promoting cause of *e* is that *c* raise *e*'s *objective* probability.

It would be easy to model cause-to-effect reasoning if this were true. As many critics have noted, however, there are cases in which the occurrence of a promoting cause can *lower* the probability of its effect by *indicating* that some more potent cause of *e* did *not* occur or that a stronger *e*-inhibitor did occur. Lewis called this type of reasoning "backtracking" and Pearl refers to it as "back-door" inference. In general, a cause *c*'s total evidential import for effect *e* is the result of two factors:

- Cause-to-effect Inference. *c* is evidence for (against) *e* in virtue of the fact that, *inter alia*, *c* promotes (inhibits) *e* more strongly than it inhibits (promotes) *e*.
- Backtracking Inference. *c* is evidence for (against) *e* in virtue of the fact that, *inter alia*, *c* indicates the presence of events or conditions (other than its own effects) that are positive (negative) indicators of *e*.

Here is an example: Suppose I encounter Jacob cramming frantically for an exam. I might believe that frantic cramming can contribute positively to his performance, and still see it as evidence *against* his doing well because of what it says about his preparation. After all, students often cram because they have not studied in advance. So, even though Jacob's frantic pre-exam cramming is a (weak) promoting cause of academic success it is strong evidence against such success in virtue of what it indicates about his study habits. In the face of counterexamples like this it is clear that promoting (inhibiting) causes need not provide evidence in favor of (against) their effects.

¹ There are Bayesian accounts of evidence that do not have this *incrementalist* character. See Joyce (2004) for relevant discussion.

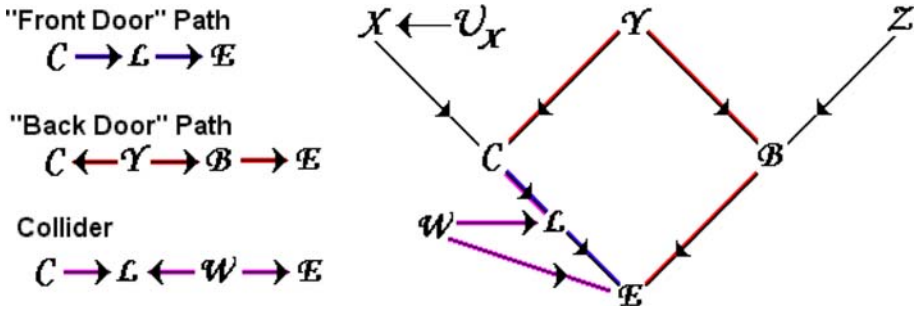
Still, that cannot be the whole story. Some suitably hedged and weakened version of the bold conjecture has *got* to be true if we are to have any hope of portraying our common patterns of reasoning as even remotely rational. So much of inductive inference involves uncritically citing causes that it would be a rather spectacular failure of collective rationality if this did not turn out to be a legitimate way of presenting evidence. My aim is to defend a weak version of the “causes are evidence” thesis by showing how it is possible, *in some cases*, to isolate a sense in which causes provide evidence for their effects in virtue of being causes. While causes do not raise the (subjective) probabilities of their effects overall, they do raise them once all the “noise” generated by backtracking inferences has been filtered out.

When asking about the nature and extent of the evidence that c provides for e we must always keep in mind that the *overall* evidential impact of c on e is a combination of the evidence that c provides for or against e *in virtue of being its cause*, and the evidence that c provides for or against e via backtracking routes. To understand how causes provide evidence about their effects we must separate these two types of evidence. Following Pearl (2000) call this the

Identification Problem. Given that a person believes c is a cause of e , can we find a way to isolate that portion of c 's evidential import for e that is due to its role as a cause of e as opposed to that portion that is due to its “backtracking” role as a mere indicator of events that are otherwise evidentially relevant to e ?

Solving this problem in complete generality involves finding a “front door” function $F(e, c)$ and a “back door” function $B(e, c)$, both dependent on the believer's subjective probability, such that $F(e, c)$ gives the increment of evidence that c provides about e in virtue of being one of its causes, while $B(e, c)$ is the increment of evidence that c provides about e in virtue of indicating other causes of e that are not themselves effects of c . Let me emphasize that the form of these functions will depend on, among other things, the subject's beliefs about causal connections involving variables other than e and c .

One useful and popular way to represent such causal beliefs by employing the formalism of “Bayesian networks” found in Pearl (2000) or Spirtes et al. (1993). In this framework, causal claims always involve reference an underlying family of *random variables*, denoted here using upper-case letters like C, E, V, X and so on. In the current context, we can think of these variables as finite partitions of propositions, so that, e.g., $C = \{c, \neg c\}$. Causal links are represented as arrows in a directed acyclic graph G at whose nodes the variables sit. For each *endogenous* variable, one with causes explicitly represented in the model, we can speak of its *causal parents*, its direct or immediate causes as given in the model. For each variable V , $par(V)$ is the set of all variables in the graph with arrows pointing directly into V . The model assumes that some fixed *functional pattern of dependence* determines the value of each variable on the basis of the values of its parents and variables not represented in the model, which are represented by catchall “error terms.” Here is a schematic example:



Each variable X is a function of its causal parents together with an unknown “error term” U_X that serves as a catchall for all relevant causal factors left out of the model.

Causal diagrams like this contain three sorts of paths. A “front-door” path follows arrows in the direction they point. These represent causal chains² leading from an initial cause through a (possibly empty) series of intervening links to some end effect. $C \rightarrow L \rightarrow E$ is a “front-door” path. A “back-door” path proceeds from an initial event up “against the arrows” to a “peak,” then down “with the arrows” along a different path to some final event. $C \leftarrow Y \rightarrow B \rightarrow E$ is a back-door path. The peak of a back-door path is a common cause of every event down either of its sides, e.g., Y is a common cause of C and E . Other paths enter some node via an incoming arrow and exit that node against an incoming arrow. This configuration is a *collider*. Path $C \rightarrow L \leftarrow W \rightarrow E$ contains a collider at L .

These three sorts of paths correspond to different ways in which information about a cause can alter the probability of an effect. Imagine a person who is certain that the causal relations are as the graph describes. Suppose also that this person is somewhat uncertain about the *values* of variables appearing in the graph, but that she has a definite subjective probability for each possible set of such values. When such a person learns the value of some variable she will update her subjective probability by Bayesian conditioning. We can view the ensuing change in the probability of values for other variables as the result of “flows of evidence,” thought of as increments or decrements of subjective probability, moving from the first variable to the second via the various paths that connect them. Evidence can only flow via front-door paths and back-door paths. Colliders are evidential dead ends; unless there are other routes between the endpoints, learning the value of a variable on one incoming path conveys nothing about the values of variables on other incoming paths.

The Identification Problem involves figuring how much of the incremental change in the person’s subjective probability for the effect event $E = e$ is due to “front-door” evidence from the cause event $C = c$, and how much is due to

² Depending on the values of the variables, these chains can describe either promoting or inhibiting causes. For example, if $C = \{c, -c\}$ and $E = \{e, -e\}$ then a chain of arrows from C to E is consistent with c promoting e and inhibiting $-e$ and with $-c$ promoting $-e$ and inhibiting e or with any other set of such causal relations.

backtracking. F is meant to measure the amount of incremental evidence that flows via front-door paths, and B is meant to measure the amount that flows along back-door paths. Thus, a first step in our problem is this: given a subjective probability P and a causal graph G , find functions $F(E, C)$ and $B(E, C)$, dependent on P and G , such that, for $E = e$ and $C = c$, $F(e, c)$ measures the evidence that “flows” from c to e along front-door paths and $B(e, c)$ measures the evidence that “flows” from c to e along back-door paths.

Proponents of the causal Bayesian network program have given us the resources to solve the Identification Problem in an important special case. I will follow Pearl’s presentation, but Spirtes, Glymour, and Scheines adopt a similar approach. A guiding assumption of both is that uncertainty about unidentified error terms *in a correctly specified causal graph* should not create correlations among variables in the model. In particular, we shall make the following two assumptions³:

- *Sufficiency*. If X and Y are correlated given a variable Z that is not among the effects of either X or Y , and neither X nor Y is an effect of the other, then there exists some variable C that is a common cause of both X and Y (so, there are front door paths from C to X and from C to Y).
- *Casual Markov*. Conditional on values for its causal parents, any variable X is independent of all others, save its effects. So, if V is the set of all X ’s causal parents, then there exist values of X , Y and V such that $P(y/x \ \& \ v) \neq P(y/v)$ only if there is a front-door path from X to Y .⁴

These require, respectively, that every correlation is explained by the existence of a common cause, and that a complete specification of the common causes of two variables will screen off any correlations between them except those generated by common effects. It follows that variables which are not connected by back-door paths or front-door paths are uncorrelated.

In such a graph the evidential import of any backtracking inference from C to E is exhausted by what it says about the C ’s immediate causes as represented in the graph. If the values of C ’s parents are known, and if E believed to be an effect of C , then the difference between E ’s prior and posterior probability will be due solely to evidence that flows from C via “front-door” paths. Likewise, if E is not among C ’s effects, so that no front-door path connects them, then any difference between $P(E/C)$ and $P(E)$ is due solely to backtracking. More generally, in a sufficient Markov graph the full difference between the prior for E prior and its posterior given C will be traceable to distinguishable flows of evidence along front-door or back-door paths *that are explicitly represented in the model*. This imposes the following constraints on any solution to the identification problem:

- (a) If E is an effect of C (i.e., there is a front-door path from C to E) and if $P(v) = 1$ for some assignment of values to C ’s parent variables, then $F(E, C) = P(E/C) - P(E)$ and $B(E, C) = 0$.

³ A third condition of “Faithfulness” is often imposed, but this will not play any role in what follows.

⁴ Here and hereafter a lower case letter appearing in some probabilistic relationship will indicate a *value* of the variable. An upper case letter means that the relation holds for *all* the variable’s values.

- (b) If E is not an effect of C (i.e., there is no front-door path from C to E), then $F(E, C) = 0$ and $B(E, C) = P(E/C) - P(E)$.
- (c) Together F and B exhaust the incremental evidential support that C provides for E , so that $P(E/C) - P(E) = F(E, C) + B(E, C)$.

To determine the form of F and B in more general contexts we can employ a trick that Pearl calls “adjustment for direct causes.” For each fixed value v of C ’s parent variables, the quantity $P(E/C \ \& \ v)$ is the subject’s probability for E upon learning C and $V = v$. Since knowledge of C ’s parents screens off backtracking routes from C to E we can compute the expected “front-door” effect, what Pearl calls the “causal effect of C on E ,” by averaging over these conditional probabilities.

Pearl’s “Adjustment for Direct Causes”. In a sufficient Markov graph, the “causal effect of c on e ” can be computed as $P(E \mid do(C) = c^*) = \sum_v P(v) \cdot P(E/c^* \ \& \ v)$ where v ranges over possible values of (all) C ’s parent variables.

I will call this *Pearl’s adjustment function*, and will read $P(E \mid do(C) = c^*)$ as “ E ’s probability adjusted for $C = c^*$.”

As Pearl tells it, E ’s probability adjusted for c is its probability after an *intervention* that “sets” C ’s value to c^* . We will discuss the idea of an intervention shortly, but the crucial point now is that such an intervention puts a person in a position to incorporate the event $C = c^*$ into her beliefs while ignoring its causal ancestry.

Pearl’s Interventionist Interpretation. $P(E \mid do(C) = c^*)$ is E ’s probability subsequent to an intervention that sets C ’s value to c in a way that makes C ’s causal history irrelevant to E ’s value.

The post-intervention probability for E thus obtained clearly reflects no “back-door” information. If we focus on this and ignore the interventionist interpretation, we can give a more “evidential” reading of Pearl’s adjustment function.

Evidential Interpretation: $P(E \mid do(C) = c^*)$ is the degree to which $P(E/c^*)$ is based on front-door inferences from the information that $C = c^*$.

Pearl’s function is close to the F we seek, but it not does measure the *incremental* impact of front-door evidence. One way to turn it into a measure of incremental support would be by subtracting E ’s unconditional probability from its probability adjusted for c^* . Unfortunately, this will not work because, for any given value e of E , $P(e)$ might already be contaminated by front-door evidence. If, say, the subject sees c^* as a promoting cause of e , then e ’s probability might be high largely because c^* ’s is high. Were we then to subtract e ’s unconditional probability from its probability adjusted for c^* we would delete much of the front-door evidence that c^* provides for e . A better strategy is to recognize that $P(E)$ itself has a front-door component given by

$$P(E \mid do(C) = \mathbf{T}) = \sum_c P(c) \cdot P(E \mid do(C) = c) = \sum_c \sum_v [P(c) \cdot P(v)] \cdot P(E/c \ \& \ v)$$

where c ranges over possible values of C and v ranges over possible values of (all) C ’s parent variables. This measures the expected degree to which the different

values of the causal variable C will affect the effect E 's probability. Given this we can define the front-door function F as the difference between E 's probability adjusted for c and E 's probability adjusted for \mathbf{T} , so that

$$\begin{aligned} F(E, c^*) &= P(E \mid do(C) = c^*) - P(E \mid do(C) = \mathbf{T}) \\ &= \sum_v \sum_c P(v) \cdot [P(E/c^* \& v) - P(c) \cdot P(E/c \& v)] \end{aligned}$$

where, again, v ranges over values of (all) C 's parent variables and c ranges over the values of C .

If this is the right form for F then, since all incremental evidence in a sufficient Markov graph either comes in through the front-door or through the back-door, we can define B by subtracting $F(E, c^*)$ from $P(E/c^*) - P(E)$. After a little calculation, this works out to

$$\begin{aligned} B(E, c^*) &= [P(E/c^*) - P(E)] - F(E, c^*) \\ &= \sum_v [P(E/c^*) - P(v)] \cdot P(E/c^* \& v) - \sum_c \sum_v [P(c \& v) - P(v) \cdot P(c)] \cdot P(E/c \& v) \end{aligned}$$

It is easy to show that these measures meet constraints (a–c) listed above. Moreover, it should be clear that if Pearl's adjustment function measures the extent to which E 's probability given c^* is based on front-door inferences from $C = c^*$, then these equations are the correct expressions for the cause-to-effect and backtracking parts of the incremental evidence that c provides for e . So, in sufficient Markov graphs, $F(E, c^*)$ and $B(E, c^*)$, respectively, give the cause-to-effect and backtracking components of $P(E/c^*) - P(E)$ on an "evidential" reading of $P(E \mid do(C) = c^*)$ and $P(E \mid do(C) = \mathbf{T})$.

There are two worries one might have about the evidential interpretation. First, portraying F and B as measures of front-door and back-door evidence flows is hopeless, even in sufficient Markov graphs, unless Pearl's function accurately measures the part of E 's posterior probability that is based on front-door inferences from $C = c$. Pearl argues that this follows from his interventionist reading of the adjustment function. If this is the only way to derive the evidential interpretation, however, then we would be committed to interventionism, which would be a problem for anyone not impressed by this conception of causation. Second, the sufficiency and Markov conditions are very strong requirements that seem essential to the definition of F and B . One might worry that insights gained in the context of sufficient Markov graphs are unlikely to carry over to other situations. I will discuss the first issue in detail here, and will have some "hand-waving" things to say about the second in my discussion of imaging.

Interventionists, like Pearl, block backtracking inferences by invoking special *intervention* events that fix values of variables in ways that circumvent their usual causes. Interventions "manipulate" variables "locally" by "setting" them to specific values *without disturbing any of the patterns of causal dependency that hold between other variables in the model*. The choice of the terms "intervention" and (even worse) "manipulation" is unfortunate, as it suggests that the relevant notions are somehow tied to human purposes and activities. As one reads the literature, however, it becomes clear that this connection to human actions and purposes is

beside the point. Indeed, many authors,⁵ including Pearl, make it clear that interventions need not be feasible. At bottom, an intervention is just a certain sequence of events that fixes a value for a variable in a particular way. When interventionists talk about “setting” a given variable to a certain value, they are imagining that the variable comes to have that value via a sequence of events that does nothing to disturb the patterns of causal dependency that hold between other variables. Since such interventions “switch off” the links between the target variable c and its immediate causes, and since backtracking inferences in sufficient Markov graphs amount to inferences about direct causes other than c , it follows that the revisions of probability needed to account for an intervention will be results of cause-to-effect inferences alone. As we have seen, the rule for calculating probabilities given an intervention is Pearl’s adjustment function.

There is a lot to like the interventionists’ approach to causation and its epistemology. Interventions do *sometimes* occur, and can sometimes be made to occur. Moreover, there is no doubt that the ability to carry out controlled experiments by “setting” values for variables is one of the most basic ways we have of learning about causal connections. (The main attraction of interventionism for statisticians, I think, is that it ties the meaning of causal claims quite closely to their method of verification.) Interventionists are also correct in thinking that we can remove back-door inferences by employing Pearl’s adjustment function.

They are wrong, however, to insist on a tight connection between rules for modifying probabilities and intervention events. It is a mistake to tie causal reasoning too closely to the possibility of interventions since it is a contingent matter whether a given causal connection admits of interventions in any realistic sense. A person who discovers that an intervention has set C to c^* must be in a position to entirely ignore C ’s usually causal history *as well as the history of the intervention event*. This is something that one is rarely in a position to do for even the most garden-variety causal claims.

Fortunately, Pearl’s rule for computing “causal effects” need not be interpreted as being *about* interventions at all. Rather than thinking of e ’s probability adjusted for c as a “post intervention” belief state, it is better seen as a sort of provisional belief revision function in which we hold the probabilities of parent variables fixed while *supposing*, in a very particular way, that a cause occurs. One can perform this operation mentally, as it were, even if one does not believe that the event being supposed can come about except in ways that reveal things about its causal ancestry. We can suppose that an event occurs in a variety of ways. One of them involves treating it as both a cause and indicator of other events. Another involves seeing it as a pure indicator while ignoring what it causes. Yet another involves viewing it only a cause and ignoring what it indicates. We employ this last, causal form of supposition in cause-to-effect inference. The point I want to emphasize is that the usefulness of Pearl’s adjustment function as a rule for causal supposition in no way depends on our being able, even in principle, to place ourselves in an epistemic situation where E ’s probability adjusted for c^* agrees with our subjective probability for E , as it would after an intervention. In some *special* situations we

⁵ Woodward (2005) is especially good on this point.

can learn that an intervention on c^* occurs, and so come to have $P(E \mid do(C) = c^*)$ as our posterior for E . However, (a) there are many other ways to arrive at this same final probability (e.g., by way of experiments involving randomized trials), and (b) even if we cannot arrive at a state in which our unconditional probability for E is equal to its adjusted probability, this does not prevent us from using the latter to understand cause-to-effect inferences. So, Pearl's rule is best seen as one of causal *supposition* that we can use to answer certain kinds of "what if" questions, whether or not any intervention occurs, or could occur.

We can buttress this point by seeing how the adjustment rule arises naturally on other metaphysical pictures of causation. Consider first a counterfactual model. Those attracted to counterfactualism will think of arrows in causal graphs as standing for patterns of counterfactual dependence. The simplest sort of theory of this sort, which is pretty much the one found in Lewis (1973) and (1979), has it that a causal arrow from C to E means that for each value c of C there is a value e^c of E such that the counterfactual conditional $c \square \rightarrow e^c$ holds when $\square \rightarrow$ is interpreted in a "non-backtracking" way. (A more sophisticated view, akin to that found in Lewis (2000), would also insist that values of c that are sufficiently alike should be associated with values of e^c that are also sufficiently alike.) On such views, the "causal effect" of c on e will be given by what Lewis calls the *image* of e on c , here be denoted as $P(EC)$. Lewis introduced imaging to explain how subjective probabilities can be assigned to counterfactual conditionals. Probabilities are thought of as attaching to "possible worlds" that bear certain (contextually defined) *similarity* relations to one another. When evaluating counterfactual claims, we judge worlds similar to the extent that they agree about the sorts of facts we deem it important to hold fixed (in a context). In the context of causal evaluations, the kind of facts that get held fixed are just the sort that figure into Skyrms' K -partitions: facts that do not depend causally on whether or not the cause occurs, but which are causally relevant to the occurrence of the effect.⁶

For purposes of defining imaging Lewis assumed that for each possible world w and each proposition c there would be another world w^c that entails c and is more *like* w than any other world that entails c . This is the "limit assumption." In imaging on c we take similarity among worlds into account by making sure that all of the probability that gets moved off each $\neg c$ -world is shifted to the unique c -world most like it. If the limit assumption holds, this process yields a probability function that accurately reflects the counterfactual supposition of $C = c$.

The problem with this definition, as Lewis was fully aware, is that there will rarely be a *unique* c -world that is *most* like any given $\neg c$ -world. Gärdenfors (1986, pp. 108–113) proposed a generalization of imaging that applies when the limit assumption fails. Gärdenfors assumed that there is always a *set* $w[c]$ of worlds in c that are most like w . In this case, he argued, the image of P on $C = c$ can be written as $P(e|c) = \sum_{z \in e} \sum_w P(w) \cdot \rho^c(w, z)$ where w and z range over possible worlds and $\rho^c(w, z)$ is a probability function for each w that is uniformly zero for z outside $w[c]$,

⁶ Lewis seeks to characterize these facts in non-causal terms. For purposes evaluating causal counterfactuals, worlds are similar when they agree about all particular facts until just before the cause occurs, and then agree afterward with respect to laws of nature. The occurrence of the cause is thus a "small miracle". See Lewis (1979) for details.

and which specifies the proportion of w 's probability that is shifted to z . Gärdenfors proves that general imaging is the only probability revision operation that sends all w 's probability to the set of c -worlds most like w and preserves mixtures.

Theorem. $P(\bullet|c)$ is a general imaging function if and only if $P = \lambda \cdot P_1 + (1 - \lambda) \cdot P_2$ implies $P(\bullet|c) = \lambda \cdot P_1(\bullet|c) + (1 - \lambda) \cdot P_2(\bullet|c)$ when $0 \leq \lambda \leq 1$.

There is, however, a hidden assumption here, which reveals something significant about general imaging. Gärdenfors's proof only goes through if the function $\rho^c(w, z)$, which determines the proportion of w 's probability that is shifted to various worlds in $w[c]$, is assumed *not* to depend on the prior distribution of probabilities over worlds in $w[c]$. This is a substantive assumption, which fails for the sort of imaging operators that are useful for representing causal knowledge.

Joyce (1999) endorsed Gärdenfors's generalization, and even suggested that $\rho^c(w, \bullet)$ should be the uniform distribution over $w[c]$. This might be called "Laplacian" imaging, since it invokes a kind of "principle of indifference". As I wrote then,

The imaging function... reflects the [believer's] judgments about similarity among possible worlds. These judgments will *not* depend on how likely she takes these worlds to be; the only place where her subjective probabilities enter the equation is through the $P(w)$ term. (1999, p. 198)

I wish I had those lines back since it now seems obvious that imaging *must* involve the *combined* effects of judgments of similarity among worlds and prior probabilities. When the information about similarity runs out, and an imager is left with a non-trivial set of c -worlds that are most like w , she still has excellent reasons for treating some worlds in this set differently from others. After all, she began by regarding some worlds in the set as more likely than others, and her evidence has not changed. Imaging should thus be *Bayesianized*, so that probabilities are spread over sets of "most similar" worlds in a way that preserves the imager's prior beliefs. More precisely, we should adopt a form of imaging that distributes w 's probability over worlds in $w[c]$ in direct proportion to their prior probability, so that $\rho^c(w, z) = P(z|w[c])$ and

$$P(e|c^*) = \sum_{z \in e} \sum_w P(w) \cdot P(z|w[c^*])$$

$$P(e|\mathbf{T}) = \sum_c P(c) \cdot [\sum_{z \in e} \sum_w P(w) \cdot P(z|w[c])].$$

It follows directly that for each value e of E

$$F(e, c^*) = P(e|c^*) - P(e|\mathbf{T}) = \sum_w \sum_c \sum_{z \in e} P(w) \cdot P(c) \cdot [(P(z|w[c^*]) - P(z|w[c])]$$

$$B(e, c^*) = [P(e/c^*) - P(e)] - F(e^*, c)$$

These formulas can be greatly simplified when the similarity relation among worlds is *stratified* for c^* in the sense that, for any worlds w_1 and w_2 , $w_1[c^*]$ and $w_2[c^*]$ are either disjoint or identical. The similarity relation then defines an equivalence \sim among worlds, where $w_1 \sim w_2$ if and only if $w_1[c^*] = w_2[c^*]$. The associated set of \sim -equivalence classes X is then a partition, and, for each cell x in X ,

every world in c^* & x is equally like any world in $\neg c^*$ & x . When the set of equivalence classes is countable, Bayesian imaging has the simple form

$$P(E \setminus c^*) = \sum_x P(x) P(E/c^* \& x)$$

and $F(E, c^*)$ and $B(E, c^*)$ assume the simple form they do in the interventionist approach. This should remind you of Skyrms' $P^c(e)$ function with the x 's playing the role of the k 's and of Pearl's adjustment formula with the x 's ranging over the partition of values of C 's causal parents. Pearl himself points out that his function can be given an imaging interpretation provided that (i) possible worlds are identified with possible assignments of values to variables in a causal graph, and (ii) worlds in which the values of C 's parent variables are the same are deemed equally similar. Proviso (i) is reasonable. If a causal graph represents all the causal structure that we take to be relevant to a problem, it makes sense to think of the relevant possible worlds as assignments of values to all the variables represented in the model. (Strictly speaking, worlds are assignments of values to all variables, whether represented in the model or not, but nothing will hang on this here.) Point (ii) is less straightforward. Variables other than C 's parents only seem irrelevant to the problem because the Markov condition holds. But, in a Bayesian context Markov describes a relationship between a graph and a person's subjective probability, and facts about similarities among worlds should not depend on subjective probabilities. Instead of (ii), worlds should be deemed similar, for purposes of computing the image of E on c^* , just in case they agree on the values of all variables that describe causal ancestors of E which are not among c^* 's effects.

When we interpret the similarity relation this way, the imaging approach has the advantage of applying when the sufficiency condition fails. Imagine a person who knows that two variables are neither causes of one another nor are effects of a common cause. Despite this known causal independence, the variables can still be correlated in the person's subjective probabilities. Berkson's paradox (1946) provides a nice illustration. Suppose that, after extensive study, it is determined that the quality of a professor's teaching evaluations are entirely determined by two factors, the professor's physical attractiveness (C) and his or her sense of humor (B). These factors are entirely uncorrelated in the populations at large and are known to be causally independent. In a crass attempt to improve teaching evaluations, a dean decides to fire everyone in the philosophy department and to replace them with witty and/or attractive people, largely ignoring factors such as philosophical knowledge or talent. After this policy has been enacted, we learn that Joe, a new hire in philosophy, is handsome c , but without learning anything about how well he tells a joke. To what extent does this support the conclusion that Joe will receive strong teaching evaluations e ?

Since we know that good looks and a sense of humor are not directly causally connected or effects of a common cause, it might seem that we can isolate the cause-to-effect increment of evidence that Joe's good looks provides for his success by setting $B(E, c) = 0$ and $F(E, c) = P(E/c) - P(E)$. There is, after all, no backdoor path from C up through some common cause and down via B to E . Surprisingly, however, there is a correlation between C and B even though no common cause explains it. As Berkson noted, if variables V_1 and V_2 are uncorrelated in the larger

population, a process that chooses individuals according to whether they possess one or both of V_1 and V_2 will, with high probability, create a subpopulation in which V_1 and V_2 are *anti*-correlated. Joe's good looks are evidence that he can't tell a joke!⁷ In light of this, the increment $P(E/c) - P(E)$ will reflect both the direct evidence that c provides as a result of being a cause of E and the "backtracking" evidence that it provides as a result of being a negative indicator of a sense of humor.

Since we lack a common cause of C and B , the interventionist strategy does not tell us what to do. Still, one can reason as follows: Assume, for simplicity, that $C = \{\text{handsome} (= c), \text{unsightly}\}$ and $B = \{\text{funny} (= b), \text{dull}\}$ are binary variables. Given that C and B are causally independent, we can define a stratified similarity relation, for purposes of imaging on C , in which, no matter how attractive Joe is, all worlds in which he is funny are regarded as maximally similar to one another, and all worlds in which he is dull are regarded as maximally similar to one another. Then,

$$P(E \setminus c) = P(b) \cdot P(E/c \ \& \ b) + P(\neg b) \cdot P(E/c \ \& \ \neg b)$$

is the non-backtracking, causal probability for the event of Joe's getting a positive teaching evaluation. The front door and backdoor components of the evidence that c provides for E then work out to be

$$F(E, c) = P(\neg C) \cdot [P(E \setminus c) - P(E \setminus \neg c)]$$

$$B(E, c) = P(\neg C) \cdot [(P(E/c) - P(E \setminus c)) - (P(E/\neg c) - P(E \setminus \neg c))]$$

These are precisely the answers one would get if one had a common cause to use as a screen (and there were only two binary variables known to be causes of C). Thus, in at least some cases, the imaging approach allows us to handle problems about which interventionists are silent.⁸

That said, the two approaches are the same in spirit. If we interpret similarity among worlds in the way advocated here—i.e., as similarity *in re* values of all variables other than c that are not "downstream" of c in a causal graph—then the image of E on c still coincides with Pearl's E adjusted for c whenever the Markov and sufficiency conditions apply. For if x ranges over all possible assignments of variables not downstream of C and if V ranges over values of C 's parents, then (since C 's parents are not downstream from C) each conjunction $x \ \& \ v$ will either be contradictory or equivalent to x . This allows us to write

⁷ If this seems counterintuitive, it helps to remind oneself that the only way that people who are not good looking can get into the department is by having a good sense of humor. Likewise, the only way people without a sense of humor can get in is by being good-looking. This explains the anti-correlation.

⁸ We can extend the approach described here to cases in which we know of a common cause Y of two correlated variables C and B , but not one that is detailed enough to fully screen-off the correlation. Here, we can block the backtracking inference by using Bayesian imaging on the partition of all conjunctions $s \ \& \ b$, for s a value of S and b a value of B . This maneuver preserves the probabilities of all the S values and the probabilities of all the B values conditional on the S values, thereby preserving the probabilities of all the B values. Once again, the backtracking inference is blocked by the imaging method.

$$P(E \setminus c) = \sum_v \sum_x P(x, \&v) \cdot P(E/c \&x \&v)$$

By Markov this is equivalent to

$$P(E \setminus c) = \sum_v \sum_x P(x \&v) \cdot P(E \setminus c \&v).$$

Finally, since X ranges over a partition we have

$$\begin{aligned} P(E \setminus c) &= \sum_v \sum_x P(x \&v) \cdot P(E/c \&v) \\ &= \sum_v P(E/c \&v) [\sum_x P(x \&v)] \\ &= P(E | do(C) = c) \end{aligned}$$

Thus, on an understanding of similarity among worlds that is stratified because worlds agreeing on all “non-downstream” causal variables are equally similar, the counterfactualist approach to causal reasoning gives us exactly the same “causal effect” function as does the interventionist approach. The front-door and back-door functions, F and B , will be the same as well.

The same is true of the broadly “probabilistic” accounts of causation defended in Skyrms (1980), Suppes (1970), and Eells (1991). These analyses yield the same results as imaging does, provided that the latter is based on (stratified) similarity relation that produces a partition that is fine enough to prevent any “hidden” backtracking, but coarse enough to reveal only genuine causal relationships. To see what I mean, imagine that the effect E is the result of two common causes, $C = \{c, \sim c\}$ and $B = \{b_1, b_2, b_3\}$, that are known not to be causally dependent on one another and not to be effects of a common cause. For the sake of definiteness, one might imagine that (i) c is the event of some drug being administered to a patient; (ii) e is an alleviation of the symptoms of a disease; and (iii) b_1, b_2, b_3 describe background conditions against which the drug will work. Perhaps b_1 says that the patient is no more than 25 years old (35% of the population); b_2 says that the patient is between 25 and 65 (40% of the population); b_3 says that the patient is over 65 (25% of the population). Suppose that random drug trials yielded results with the following frequencies:

	$c \& e$	$\sim c \& e$	$c \& \sim e$	$\sim c \& \sim e$
b_1	5%	10%	15%	5%
b_2	25%	5%	5%	5%
b_3	4%	6%	6%	9%

A geriatrician might partition the background conditions into two cells $b_1 \vee b_2$ (the non-elderly) and b_3 (the elderly). Relative to this partition the drug appears causally inert both on the elderly and the non-elderly, since $P(e/c \& b_3) = P(e/b_3) = 0.4$ and $P(e/c \& (b_1 \vee b_2)) = P(e/b_1 \vee b_2) = 0.6$. On the other hand, a general practitioner might focus on the finer partition $\{b_1, b_2, b_3\}$. Relative to this division of the background conditions, the drug still has no effect on the elderly, but it

appears to promote recovery among people in their middle years $P(e/c \ \& \ b_2) = 0.833 > P(e/b_2) = 0.75$, but to prevent recovery among the young $P(e/c \ \& \ b_1) = 0.25 < P(e/b_1) = 0.5$. Whose interpretation is correct?

The answer cannot be given a priori. Moreover, whatever it is, the answer will make essential use of causal notions. The crux is whether the difference between youth versus middle age really is a causally relevant factor *in re* the drug's effects, or whether evidence that *C* provides for/against *E* in the event of youth/middle age reflects a "spurious" non-causal correlation. Suppose the drug works by regulating the production of some protein, fixing its presence in the blood at a specific level. Imagine as well that the symptoms of the disease are worse the more of the protein there is in the blood. Then, if those under 25 tend naturally to produce the protein at lower than the critical level, if those between 25 and 65 produce it at greater than the critical level, and if those over 65 tend to produce it at near the critical level, then the patient's youth versus middle-age status will make a difference to the causal impact of *c* on *e*. In this case, the right partition would be the finer one $\{b_1, b_2, b_3\}$. On the other hand, if the mechanism by which the drug works is indifferent to factors tied to differences between patients in their youth and those in their middle years, then the coarser $\{b_1 \vee b_2, b_3\}$ is preferable. Either way the correct partition should reflect all of those factors that really do make a difference to whether *C* is a cause of *E*. Only when *F* and *B* are computed relative to such a partition will they yield accurate values for the front-door and back-door components of the evidence that *C* provides for *E*. So, in addition to being equivalent to the causal graph account, the imaging approach to representing causal reasoning is also equivalent to the position defended in Skyrms (1980) provided that the similarity relation reflects all and only those factors that might make a genuine causal difference to *E* given a value for *C*.⁹

The general moral is straightforward: Whether we use causal graphs, imaging, or a probabilistic approach to causation, if we describe the essential features of the causal situation correctly, in probabilistic terms, then the incremental evidence that a cause provides for its effect in virtue of being its cause will always work out to have the same value. When all is said and done, then, the front-door and back-door functions *F* and *B* are the common property of us all, no matter how much we might disagree about the deep metaphysics of causation or about the surface formulation of causal reasoning.

References

- Berkson, J. (1946). Limitations of the application of fourfold tables to hospital data. *Biometrics Bulletin (now Biometrics)*, 2(3), 47–53.
- Eells, E. (1991). *Probabilistic causality*. Cambridge, UK: Cambridge University Press.
- Gärdenfors, P. (1982). Imaging and conditionalization. *Journal of Philosophy*, 79, 747–760.

⁹ I do not mean to suggest that the equivalence of these three approaches has gone unnoticed. In fact, I think most experts in the field hold this view. The goal, rather, is to emphasize that all three approaches to causal reasoning produce exactly the same *F* and *B* functions, and thus the same basic epistemology of cause-to-effect inference.

- Good, I. J. (1961). A causal calculus I. *British Journal for the Philosophy of Science*, 11, 305–318.
- Good, I. J. (1962). A causal calculus II. *British Journal for the Philosophy of Science*, 12, 43–51.
- Joyce, J. (1999). *The foundations of causal decision theory*. Cambridge, UK: Cambridge University Press.
- Joyce, J. (2004). Bayesianism. In A. Mele & P. Rawling (Eds.), *The Oxford handbook of rationality* (pp. 132–155). Oxford: Oxford University Press.
- Lewis, D. (1973). Causation. *Journal of Philosophy*, 70, 556–567.
- Lewis, D. (1979). Counterfactuals dependence and time's arrow". *Nous*, 13, 455–476.
- Lewis, D. (2000). Causation as influence. *Journal of Philosophy*, 97, 182–197.
- Pearl, J. (2000). *Causality*. Cambridge, UK: Cambridge University Press.
- Reichenbach, H. (1956). *The direction of time*. Berkeley and Los Angeles: University of California Press.
- Skyrms, B. (1980). *Causal necessity*. New Haven: Yale University Press.
- Skyrms, B. (1984). *Pragmatics and empiricism*. New Haven: Yale University Press.
- Spirtes, P., Glymour, C., & Scheines, R. (1993). *Causation*. New York: Springer-Verlag, prediction and search.
- Suppes, P. (1970). *A probabilistic theory of causality*. Amsterdam: North-Holland Publishing Company.
- Woodward, J. (2005). *Making things happen: A theory of causal explanation*. Oxford: Oxford University Press.