Causation[1]

Richard Scheines

Department of Philosophy
Carnegie Mellon University

1 Introduction

Practically, causation matters. Juries must decide, for example, whether a pregnant mother's refusal to give birth by caesarean section was the cause of one of her twins death. Policy makers must decide whether violence on TV causes violence in life. Neither question can be coherently debated without some theory of causation. Fortunately (or not, depending on where one sits), a virtual plethora of theories of causation have been championed in the third of a century between 1970 and 2004.

Before we sketch a few of the major ones, however, consider what we might want out of a theory of causation. First, although we can all agree that causation is a relation, what are the relata? Are causes and effects *objects*, like moving billiard balls? Are they *particular events*, like the Titanic hitting an iceberg in 1912? Or are they *kinds of events*, like smoking cigarettes and getting lung cancer? As it turns out, trying to understand causation as a relation between particular objects or events is quite a different task than trying to understand it as relation between *kinds* of occurrences or events (See sidebar 1).

Second, we want a theory to clarify, explain, or illuminate those properties of causation we can agree are central. For example, whatever causation is, it has a direction. Warm weather causes people to wear lighter clothing, but wearing lighter clothing does not cause warm weather. A theory that fails to capture the *asymmetry of causation* will be unsatisfying.

Third, we know that in many cases one thing can occur regularly *before* another, and thus appear to be related as cause and effect, but are in fact effects of a common cause, a phenomenon we will call *spurious causation*. For example, flashes of lightning appear just before and seem to cause the thunderclaps that follow them, but in reality both are effects of a common cause: the superheating of air molecules from the massive static electric discharge between the earth and the atmosphere. A good theory of causation ought to successfully separate cases of real from spurious causation.

---

[1] (in press: *New Dictionary of the History of Ideas*, Charles Scribner and Sons)

The history of thinking on causation from 1970-2004 can be organized in many ways, but the one that separates matters best, both temporally and conceptually is captured eloquently by Clark Glymour:

> Philosophical theories come chiefly in two flavors, Socratic and Euclidean. Socratic philosophical theories, whose paradigm is *The Meno*, advance an analysis (sometimes called an "explication"), a set of purportedly necessary and sufficient conditions for some concept, giving its meaning; in justification they consider examples, putative counterexamples, alternative analyses and relations to other concepts. Euclidean philosophical theories, whose paradigm is *The Elements*, advance assumptions, considerations taken to warrant them, and investigate the consequences of the assumptions. Socratic theories have the form of definitions. Analyses of "virtue," "cause," "knowledge," "confirmation," "explanation," are ancient and recent examples. Euclidian theories have the form of formal or informal axiomatic systems and are often essentially mathematical: Euclid's geometry, Frege's logic, Kolmogorov's probabilities, … . That of course does not mean that Euclidean theories do not also contain definitions, but their definitions are not philosophical analyses of concepts. Nor does it mean that the work of Euclidean theories is confined to theorem proving: axioms may be reformulated to separate and illuminate their contents or to provide justifications. (Glymour, 2004).

For causation, Socratic style analyses dominated from approximately 1970 to the mid 1980s. By then, it had become apparent that all such these theories either invoked non-causal primitives that were more metaphysically mysterious than causation itself, or were circular, or were simply unable to account for the asymmetry of causation or separate spurious from real causation. Slowly, Euclidean style theories replaced Socratic ones, and by the early 1990s a rich axiomatic theory of causation had emerged that combined insights from statisticians, computer scientists, philosophers, economists, psychologists, social scientists, biologists, and even epidemiologists.

2 The 1970s and Early 1980s: The Age of Causal Analyses

2.1 The Counterfactual Theory

In the late 1960s, Robert Stalnaker began the rigorous study of sentences that assert what are called contrary to fact conditionals. For example, "If the Sept. 11th, 2001 terrorist attacks on the U.S. had not happened, then the U.S. would not have invaded Afghanistan shortly thereafter." In his classic 1973 book *Counterfactuals*, the late David Lewis produced what has become the most popular account of such statements. Lewis' theory rests on two ideas: the existence of alternative "possible worlds," and a similarity metric over these worlds. For example, it is intuitive that the possible world identical to

our own in all details except for the spelling of my wife's middle name ("Anne" instead of "Ann") is closer to the actual world than one in which the asteroid that killed the dinosaurs missed the earth and primates never evolved from mammals.

For Lewis, the meaning and truth of counterfactuals depend on our similarity metric over possible worlds. When we say "if A hadn't happened, then B wouldn't have happened either," we mean that for each possible world W_1 in which A didn't happen and B did happen, there is at least one world W_2 in which A didn't happen and B didn't happen that is *closer* to the actual world than W_1. Lewis represents counterfactual dependence with the symbol: $\Box\!\!\rightarrow$, so P $\Box\!\!\rightarrow$ Q means that, among all the worlds in which P happens, there is a world in which Q also happens that is closer to the actual world than all the worlds in which Q doesn't.

That there is some connection between counterfactuals and causation seems obvious. We see one event A followed by another B. What do we mean when we say A caused B? We might well mean that if A hadn't happened, then B wouldn't have either. If the Titanic hadn't hit an iceberg, it wouldn't have sunk. Formalizing this intuition in 1973, Lewis analyzed causation as a relation between two events A and B that both occurred such that two counterfactuals hold:

1. A $\Box\!\!\rightarrow$ B, and
2. ~A $\Box\!\!\rightarrow$ ~B

Because A and B both already occurred, 1 is trivially true, so we need only assess 2 in order to assess whether A caused B.

Is this analysis satisfactory? Even if possible worlds and a similarity metric among them are clearer and less metaphysically mysterious than causal claims, which many dispute, there are two major problems with this account of causation. First, in its original version it just misses cases of overdetermination or pre-emption, that is, cases in which more than one cause was present and could in fact have produced the effect (see sidebar 2).[2]

Even more importantly, Lewis' counterfactual theory has a very hard time with the asymmetry of causality and only a slightly better time with the problem of spurious causation. Consider a man George who jumps off the Brooklyn Bridge and plunges into the East River.[3] On Lewis' theory, it is clear that it was jumping that caused George to plunge into the river, because had George not jumped, the world in which he didn't plunge is closer to the actual one than any in which he just happened to plunge for some other reason at approximately the same time. Fair enough. But consider the opposite

[2] Lewis (2000) and many others have amended the counterfactual account of causation to handle problems of overdetermination and pre-emption, but neither his nor others have yet satisfactorily handled the asymmetry of causality.

[3] This example is originally from Horacio Arlo-Costa, and discussed in Hausman, 1998, pp. 116-117.

direction: if George hadn't plunged, then he wouldn't have jumped. Should we assent to this counterfactual? Is a world in which George didn't plunge into the river and didn't jump closer to the real one than any in which he didn't plunge but *did* jump? Most everyone except Lewis and his followers would say yes. Thus on Lewis' account jumping off the bridge caused George to plunge into the river, but plunging into the river[4] also caused George to jump.

For the problem of spurious causation, consider Johnny, who gets infected with the measles virus, runs a fever and shortly thereafter gets a rash. Is it reasonable to assert that if Johnny had not gotten a fever, he would not have gotten a rash? Yes, but it was not the fever that caused the rash, it was the measles virus. Lewis responded to this problem by prohibiting "backtracking" (Lewis, 1986), and to the problem of overdetermination and pre-emption with an analysis of "influence" (Lewis, 2000), but the details are beyond our scope.

2.2 Mackie's Regularity Account

Where David Lewis tried to base causation on counterfactuals, John Mackie tried to extend Hume's idea that causes and effects are "constantly conjoined," and use the logical idea of necessary and sufficient conditions to make things clear. In 1974, Mackie developed an analysis of causation in some part aimed at solving the problems that plagued Lewis' counterfactual analysis, namely overdetermination and pre-emption. Mackie realized that *many* factors combine to produce an effect, and it is only our idiosyncratic sense of what is "normal" that draws our attention to one particular feature of the situation, such as hitting the iceberg. It is a set of factors, e.g., A: air with sufficient oxygen, B: a dry pile of combustible newspaper and kindling, and C: a lit match that *combine* to cause D: a fire. Together the *set* of factors A, B, and C are *sufficient* for D, but there might be other sets that would work just as well, for example A, B, and F: a bolt of lightning. If there was a fire caused by a lit match, but a bolt of lightning occurred that also would have started the fire, then Lewis' account has trouble saying that the lit match caused the fire, because the fire would have started without the lit match, or put another way, the match wasn't necessary for starting the fire. Mackie embraces this idea, and says that X is a cause of Y just in case X is an Insufficient but Necessary part of an Unnecessary but Sufficient *set* of conditions for Y, that is, an INUS condition. The set of conditions that produced Y need not be the only sufficient set, thus the set isn't necessary, but X should be an essential part of a set that is sufficient for Y.

Again, however, the asymmetry of causality and the problem of spurious causation wreak havoc with Mackie's INUS account of causation. Before penicillin, approximately ten percent of those people who contracted syphilis eventually got a debilitating disease called paresis, and nothing doctors could measure seemed to tell them

[4] As distinct from the *idea* or *goal* of plunging into the river.

anything about which syphilitics got paresis and which did not. As far as we know, paresis can result only from syphilis, so having paresis is by itself sufficient for having syphilis. Consider applying Mackie's account to this case. P paresis *is* an INUS condition for of syphilis, because it is sufficient by itself for having syphilis, but it is surely not a cause of it.

Consider the measles. If we suppose that when people are infected, they either show both symptoms (the fever and rash) or their immune system controls it and they show neither, then the INUS theory gets things wrong. The fever is a necessary part of a set that is sufficient for the rash: {fever, infected with measles virus}, and for that matter the rash is a necessary part of a set that is sufficient for fever: {rash, infected with measles virus}. So, unfortunately, on this analysis fever is an INUS cause of rash and rash is also a cause of fever.

2.3 Probabilistic Causality

Twentieth century physics has had a profound effect on a wide range of ideas, including theories of causation. In the years between about 1930 and 1970, the astounding and unabated success of quantum mechanics forced most physicists to accept the idea that, at bedrock, the material universe unfolds probabilistically. Past states of sub-atomic particles, no matter how finely described, do not determine their future states, they merely determine the probability of such future states. Embracing this brave new world in 1970, Patrick Suppes published a theory of causality that attempted to reduce causation to probability. Whereas electrons have only a propensity, that is, an objective physical probability to be measured at a particular location at a particular time, perhaps macroscopic events like getting lung cancer have only a probability as well. We observe that some events seem to quite dramatically change the probability of other events, however, so perhaps causes *change the probability* of their effects. If Pr(E), the probability of an event E, changes after we are told that another event C has occurred, notated Pr(E | C), then we say E and C are *associated*. If not, then we say they are *independent*. Suppes was quite familiar with the problem of asymmetry, and he was well aware that association and independence are perfectly symmetric, that is, $Pr(E) = Pr(E \mid C) \Leftrightarrow Pr(C) = Pr(C \mid E)$. He was also familiar with the problem of spurious causation, and knew that two effects of a common cause could appear associated. To handle asymmetry and spurious causation, he used time and the idea of *conditional* independence. His theory of probabilistic causation is simple and elegant:

1. C is a *prima facie* cause of E if C occurs before E in time, and C and E are associated, i.e., Pr(E) < Pr(E | C).
2. C is a *genuine cause* of E if C is a *prima facie* cause of E, and there is *no* event Z prior to C such that C and E are independent conditional on Z, i.e., there is *no* Z such that Pr(E | Z) = Pr(E | Z, C).

Without doubt, the idea of handling the problem of spurious causation by looking for other events Z that *screen off* C and E, although anticipated by Hans Reichenbach, I. J. Good and others, was a real breakthrough, and remains today a key feature of any metaphysical or epistemological account that connects causation to probability. Many other writers have elaborated a probabilistic theory of causation with metaphysical aspirations, e.g.,. Ellery Eells, David Papineau, Brian Skyrms, and Wolfgang Spohn.

Probabilistic accounts have drawn criticism on several fronts. First, defining causation in terms of probability just replaces one mystery with another. Although we have managed to produce a mathematically rigorous theory of probability, the core of which is now widely accepted, we have not managed to produce a reductive metaphysics of probability. It is still as much a mystery as causation. Second, there is something unsatisfying about using time explicitly to handle the asymmetry of causation and at least part of the problem of spurious causation (we can only screen off spurious causes with a Z that is prior in time to C).

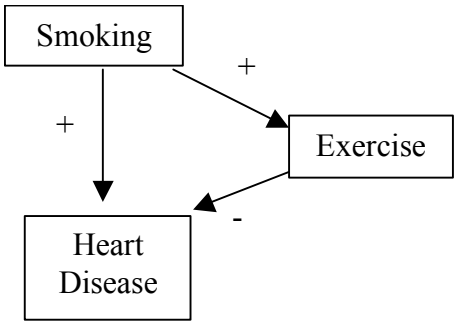

Figure 1: Cartwright's Counterexample

Third, as Nancy Cartwright persuasively argued in 1979, we cannot define causation with probabilities alone, we need causal concepts in the definiens as well as the definiendum. Consider her famous (even if implausible) hypothetical example, shown in Figure 1: smoking might cause more heart disease, but it might also cause exercise, which in turn might cause *less* heart disease. If the negative effect of exercise on heart disease is stronger than the positive effect of smoking, and the association between smoking and exercise is high enough, then the probability of heart disease given smoking could be *lower* than the probability of heart disease given not smoking, making it appear as if smoking prevents heart disease instead of causing it.

The two effects could also exactly cancel, making smoking and heart disease look independent. Cartwright's solution is to look at the relationship between Smoking and heart disease within groups that are doing the same amount of exercise, that is, to look at the relationship between smoking and heart disease *conditional* on exercise, even though exercise does not in this example come before smoking as Suppes insists it should. Why doesn't Suppes allow Zs that are prior to E but after C in time? Because that would allow situations in which although C really does cause E, its influence was entirely mediated by Z, and by conditioning on Z it appears as if C is *not* a genuine cause of E, even though it is (Figure 2).

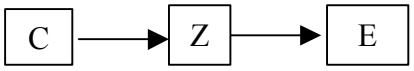

Figure 2: Z Mediates the relation between C and E

In Cartwright's language: Smoking should increase the probability of heart disease in *all causally homogenous situations* for heart disease. The problem is circularity. By referring to the causally homogenous situations we invoke causation in our definition. The moral Cartwright drew and one that is now widely accepted, is that causation is *connected* to probability, but cannot be *defined* in terms of it.

2.4 Salmon's Physical Process Theory

A wholly different account of causation comes from Wes Salmon, one of the pre-eminent philosophers of science in the later half of the 20th century. In the 1970s, Salmon developed a theory of scientific explanation that foundered partly on an asymmetry very similar to the asymmetry of causation. Realizing that causes explain their effects but not vice versa, Salmon made the connection between explanation and causation explicit. He then went on to characterize causation as an interaction between two *physical processes,* not a probabilistic or logical or counterfactual relationship between events. A *causal interaction*, according to Salmon, is the intersection of two *causal processes* and the exchange of some invariant quantity, like momentum. For example, two pool balls that collide each change directions (and perhaps speed), but their total momentum after the collision is (ideally) no different than before. An interaction has taken place, but momentum is conserved. Explaining the features of a causal process is beyond the scope of such a short review article, but Phil Dowe has made them quite accessible and extremely clear in a 2000 review article in the *British Journal for Philosophy of Science*.

It turns out to be very difficult to distinguish real causal processes from psuedo-processes, but even accepting Salmon's and Dowe's criteria, the theory uses time to handle the asymmetry of causation and has big trouble with the problem of spurious causation. Again, see Dowe's excellent review article for details.

2.5 Manipulability Theories

Perhaps the most tempting strategy for understanding causation is to conceive of it as how the world responds to an intervention, or manipulation. Consider a well-insulated, closed room containing two people. The room is 58 degrees Fahrenheit, and each person has a sweater on. Later, the room is 78 degrees Fahrenheit and each person has taken their sweater off. If we ask whether it was the raise in room temperature that caused the people to peel off their sweaters, or the peeling off of sweaters that caused the room temperature to rise, then unless there was some strange signal between the taking off of sweaters and turning up a thermostat somewhere, the answer is obvious. Manipulating the room temperature from 58 to 78 degrees will cause people to take off their sweaters, but manipulating them to take off their sweaters will not make the room heat up.

In general, causes can be used to control their effects but effects cannot be used to control their causes. Further, there is an invariance between a cause and its effects that does not hold between an effect and its causes. It doesn't seem to matter *how* we change the temperature in the room from 58 to 78 degrees or from 78 to 58, the co-occurrence between room temperature and sweaters remains. When the room is 58, people have sweaters on. When the room is 78, they don't. The opposite is not true for the relationship between the effect and its causes. It *does* matter how they come to have their sweaters on. If we let them decide for themselves naturally, then the co-occurrence between sweaters and temperature will remain, but if we intervene to make them take their sweaters off or put them on, then we will annihilate any co-occurrence between wearing sweaters and the room temperature, precisely because the room temperature will not *respond* to whether or not people are wearing sweaters. Thus, manipulability accounts nail the asymmetry problem.

They do the same for the problem of spurious causation. Tar-stained fingers and lung cancer are both effects of a common cause – smoking. Intervening to remove the stains from one's fingers will not in any way change the probability of getting lung cancer, however.

The philosophical problem with manipulability accounts is circularity, for what is it to "intervene" and "manipulate" other than to "cause." Intervening to set the thermostat to 78 is just to cause it to be set at 78. Manipulation is causation, so defining causation in terms of manipulation is, at least on the surface of it, circular.

Perhaps we can escape from this circularity by separating human actions from natural ones. Perhaps forming an intention and then acting to execute it *is* special, and *could* be used as a non-causal primitive in a reductive theory of causation. Writers like von Wright and Mackenzie have pursued this line. Others, like Paul Holland, have gone so far as to say that we have no causation without human manipulation. But is this reasonable or desirable? Virtually all physicists would agree that it is the moon's gravity that causes

the tides. Yet we cannot manipulate the moon's position or its gravity. Are we to abandon all instances of causation where human manipulation was not involved? If a painting falls off the wall and hits the thermostat, bumping it up from 58 to 78 degrees, and a half hour later sweaters come off, are we satisfied saying that the sequence: thermostat goes up → room temperature goes up → sweaters come off was not causal?

Because they failed as reductive theories of causation, manipulability theories drew much less attention then perhaps they should have. As Jim Woodward (2003) elegantly puts it:

> Philosophical discussion has been unsympathetic to manipulability theories: it is claimed both that they are unilluminatingly circular and that they lead to an implausibly anthropocentric and subjectivist conception of causation. This negative assessment among philosophers contrasts sharply with the widespread view among statisticians, theorists of experimental design, and many social and natural scientists that an appreciation of the connection between causation and manipulation can play an important role in clarifying the meaning of causal claims and understanding their distinctive features. (Woodward, 2003, p. 25).

3 The Axiomatic and Epistemological Turn: 1985-2004

Although there will always be those unwilling to give up on a reductive analysis of causation, by the mid 1980s it was reasonably clear that such an account was not forthcoming. What has emerged as an alternative, however is a rich axiomatic theory that clarifies the role of manipulation in much the way Woodward wants and connects rather than reduces causation to probabilistic independence, as Nancy Cartwright insisted. The modern theory of causation is truly interdisciplinary, and fundamentally epistemological in focus. That is, it allows a rigorous and systematic investigation of what can and cannot be learned about causation from statistical evidence. Its intellectual beginnings go back at least eighty years.

Path Analysis

Sometime around 1920, the brilliant geneticist Sewall Wright realized that standard statistical tools were too thin to represent the causal mechanisms he wanted to model. He invented "path analysis" to fill the gap. Path analytic models are causal graphs (like those shown in Figure 1 and Figure 2) that quantify the strength of each arrow, or direct cause, which allowed Wright to quantify and estimate from data the relative strength of two or more mechanisms by which one quantity might affect another. By mid-century, prominent economists (e.g. Herbert Simon and Herman Wold), and sociologists (e.g., Hubert Blalock and Otis Dudley Duncan) had adopted this representation. In several instances they made important contributions, either by expanding the representational

power of path models, or by articulating how one might distinguish one causal model from another with statistical evidence.

Path models, however, did nothing much to help model the asymmetry of causation.

Statistical Model **Path Diagram**

$Y = \beta X + \varepsilon$
$\varepsilon \sim N(0, \sigma^2)$

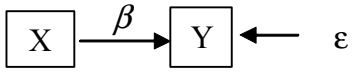

Figure 3: Path Analytic Model of X → Y

In the simplest possible path model representing that X is a cause of Y (Figure 3), we write Y as a linear function of X and an "error" term ε that represents all other unobserved causes of Y besides X. The real-valued coefficient β quantifies X's effect on Y. Nothing save convention, however, prevents us from inverting the equation and rewriting the statistical model as:

$$X = \alpha Y + \delta, \qquad \text{where } \alpha = 1/\beta \text{ and } \delta = -1/\beta\, \varepsilon$$

This algebraically equivalent model makes it appear as if Y is the cause of X instead of vice versa. Equations are symmetric, but causation is not.

Philosophy

In the early 1980s, two philosophers of causation, David Papineau and Dan Hausman, paying no real attention to path analysis, nevertheless provided major insights into how to incorporate causal asymmetry into path models and probabilistic accounts of causation. Papineau, in a 1985 article titled "Causal Asymmetry" considered the difference between 1) two effects of a common cause and 2) two causes of a common effect (Figure 4).

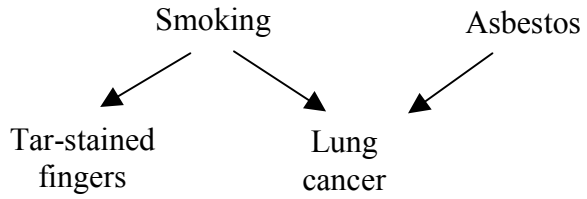

Figure 4: The Asymmetry of Common Cause and Common Effect

He argued that two effects of a common cause (tar-stained fingers and lung cancer) are *associated* in virtue of having a common cause (smoking), but that two causes of a common effect (smoking and asbestos) are not associated in virtue of having a common effect (lung cancer). In fact, he could have argued that the two effects of a common cause C *are* associated in virtue of C, but are *independent* conditional on C, whereas the two causes of a common effect E *are not* associated in virtue of E, but *are associated* conditional on E.

Dan Hausman, in a 1984 article (and more fully in a 1998 book *Causal Asymmetries*) generalized this insight still further by developing a theory of causal asymmetry based on "causal connection." X and Y are *causally connected* if and only if X is a cause of Y, Y a cause of X, or there is some common cause of both X and Y. Hausman connects causation to probability by assuming that two quantities are associated if they are causally connected, and independent if they are not. How does he get the asymmetry of causation? By showing that when X is a cause of Y, anything else causally connected to X is also connected to Y, but not vice versa (see sidebar 3).

Papineau and Hausman handle the asymmetry of causation by considering not just the relationship between the cause and effect, but rather by considering the way a cause and effect relate to other quantities in an expanded system. How does this help locate the asymmetry in the path analytic representation of causation? First, consider the apparent symmetry in the statistical model in Figure 3. X and ε are not causally connected, and have Y as a common effect. Thus following both Papineau and Hausman, we will assume that X and ε are independent, and that in any path model properly representing a direct causal relation C $\rightarrow$ E, C and the error term for E will be independent. But now consider the equation $X = \alpha Y + \delta$, which we used to make it appear that Y$\rightarrow$ X. Because of the way δ is defined, Y and δ will be associated, except for extremely rare cases.

Statistics and Computer Science

Path analytic models have two parts, a path diagram and a statistical model (see Figure 3). A path diagram is just a directed graph, a mathematical object very familiar to computer scientists and somewhat familiar to statisticians. As we have seen, association and independence are intimately connected to causation, and they happen to be one of the fundamental topics in probability and statistics.

Paying little attention to causation, in the 1970s and early 1980s statisticians Phil Dawid, David Spiegelhalter, Nanny Wermuth, David Cox, Stefan Laurizten and others developed a branch of statistics called *graphical models* that represented the independence relationships among a set of random variables with undirected and directed graphs. Computer scientists interested in studying how robots might learn began to use graphical models to represent and selectively update their uncertainty about the world, especially Judea Pearl and his colleagues at UCLA. By the late 1980s, Pearl had developed a very powerful theory of reasoning with uncertainty using Bayes Networks and the Directed Acyclic Graphs (DAGs) attached to them. Although in 1988 he eschewed interpreting Bayes Networks causally, Pearl made a major epistemological breakthrough by beginning the study of indistinguishability. He and Thomas Vermacharacterized when two Bayes Networks with different DAGs entail the same independencies, and are thus empirically indistinguishable on evidence consisting of independence relations.

Philosophy Again

In the mid 1980s, Peter Spirtes, Clark Glymour, and Richard Scheines (SGS hereafter), philosophers working at Carnegie Mellon, recognized that path analysis was a special case of Pearl's theory of DAGs. Following Hausman, Papineau, Cartwright and others trying to connect rather than reduce causation to probabilistic independence, they explicitly axiomatized the connection between causation and probabilistic independence in accord with Pearl's theory and work by statisticians Kiiveri and Speed. Their theory of causation is explicitly non-reductionist. Instead of trying to define causation in terms of probability, counterfactuals, or some other relation, they are intentionally agnostic about the metaphysics of the subject. Instead, their focus is on the epistemology of causation, in particular on exploring what can and cannot be learned about causal structure from statistics concerning independence and association. SGS formulate several axioms connecting causal structure to probability, but one is central:

> **Causal Markov Axiom**: Every variable is probabilistically independent of all of its non-effects (direct or indirect), conditional on its immediate causes.

The axiom has been the source of a vigorous debate,[5] but it is only half of the SGS theory. The second half involves explicitly modeling the idea of a manipulation, or intervention. All manipulability theories conceive of interventions as coming from outside the system. SGS model an intervention by adding a new variable external to the system which:

1. is a direct cause of exactly the variable it targets, and
2. the effect of no variable in the system,

and by assuming that the resulting system still satisfies the Causal Markov axiom.

If the intervention completely determines the variable it targets, then the intervention is *ideal*. Since an ideal intervention determines its target and thus overrides any influence the variable might have gotten from its other causes, SGS model the intervened system by "x-ing out" the arrows into the variable ideally intervened upon. In Figure 5-a for example, we show the causal graph relating room temperature and wearing sweaters. In Figure 5-b, we show the system in which we have intervened upon room temperature with I_1, and in Figure 5-c, the system after an ideal intervention I_2 on sweaters On.

[5] See the British Journal for the Philosophy of Science between 1999 and 2002.

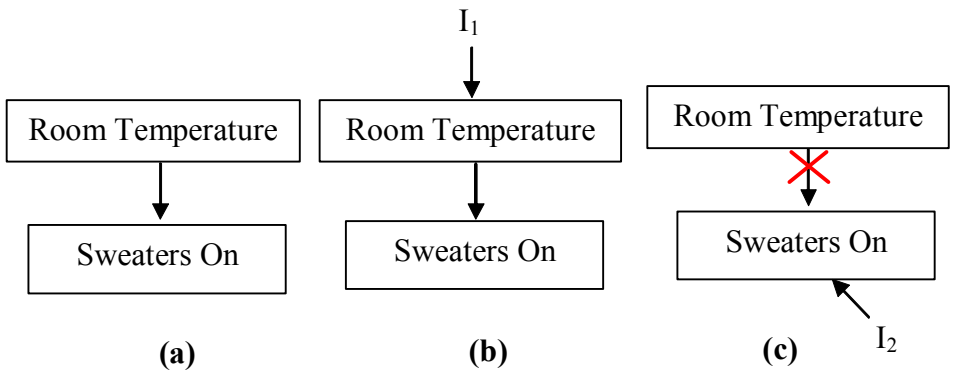

Figure 5: Ideal Interventions in SGS Theory

This basic perspective on causation, elaborated powerfully and presented elegantly by Judea Pearl (2000), has also been adopted by other prominent computer scientists (e.g., David Heckerman and Greg Cooper), psychologists (Alison Gopnik and Patricia Cheng), economists (e.g., David Bessler, Clive Granger, and Kevin Hoover), epidemiologists (Sander Greenland and Jamie Robins, biologists (e.g., William Shipley), statisticians (e.g., Stefan Lauritzen, Thomas Richardson and Larry Wasserman), and philosophers (e.g., Jim Woodward and Dan Hausman).

How is the theory epistemological? Researchers have been able to characterize precisely, for many different sets of assumptions above and beyond the Causal Markov axiom, the class of causal systems that are empirically indistinguishable, and they have also been able to automate discovery procedures that can efficiently search for such indistinguishable classes of models, including models with hidden common causes. Even in such cases, we can still sometimes tell just from the independencies and associations among the variables measured that one variable is not a cause of another, that two variables are effects of an unmeasured common cause, or that one variable is a definite cause of another. We even have an algorithm for deciding, from data and the class of models that are indistinguishable on these data, when the effect of an intervention *can* be predicted and when it cannot. For a compendium of these results and dozens of applications to real data, see (Spirtes, Glymour and Scheines, 2000; Pearl, 2000; Glymour and Cooper, 1999).

Like any new theory in town, the theory has its detractors. Philosopher Nancy Cartwright, although having herself contributed heavily to the axiomatic theory, is the most vocal recent critic of its core axiom, the Causal Markov axiom. Cartwright maintains that common causes do not always screen off their effects. Her chief counterexample involves a chemical factory, but the example is formally identical to another that is easier to understand. Consider a TV with a balky on/off switch. When turned to "on," the switch doesn't always make the picture and sound come on, but

whenever it makes the sound come on, it also makes the picture come on (Figure 6). The problem is this: knowing the state of the switch doesn't make the sound and the picture independent. Even having been told that the switch is on, for example, also being told that the sound is on adds information about whether the picture is also on.

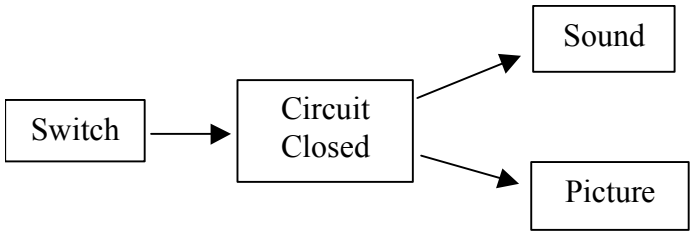

Figure 6: Non-screening off

The response of SGS and many others (e.g., Hausman and Woodward), is that it only appears as if we do not have screening off because we are not conditioning on *all* the common causes, especially those more proximate to the effects in question. They argue that we must condition on the Circuit Closed, and not just on the Switch, in order to screen off Sound and Picture.

A deeper puzzle along these same lines arises from quantum mechanics. In a famous thought experiment, Einstein, Podolosky, and Rosen consider a coupled system of quantum particles that are separated gently and allowed to diverge. Each particle is in superposition, that is, it *has no definite spin until it is measured*. Bell's famous inequality shows that no matter how far apart we allow them to drift, the measurements on one particle will be highly correlated with the other, even after we condition on the state of the original coupled system. There are no extra hidden variables (common causes) we could introduce to screen off the measurements of the distant particles. Although the details are quite important and nothing if not controversial, it looks as if the Causal Markov axiom might not hold in quantum mechanical systems. Why it should hold in macroscopic systems when it might not hold for their constituents is a mystery.

The SGS model of an intervention incorporates many controversial assumptions. In a recent tour-de-force, however, Jim Woodward (2003) works through all the philosophical reasons why the basic model of intervention adopted by the interdisciplinary view is reasonable. For example, Woodward considers why a manipulation must be modeled as a direct cause of only the variable it targets. Not just *any* manipulation of our roomful of sweater wearing people will settle the question of whether sweater wearing causes the room temperature. If we make people take off their sweater by blowing super hot air on them - sufficient to also heat the room - then we have not *independently* manipulated just the sweaters. Similarly, if we are testing to see if confidence improves athletic performance, we can't intervene to improve confidence with a muscle relaxer that also reduces motor coordination. These manipulations are "fat hand" - they directly cause more than they should.

Woodward covers many issues like this one, and develops a rich philosophical theory of intervention that is not reductive but is illuminating and rigorously connects the wide range of ideas that have been associated with causation. For example, the idea of an independent manipulation illuminates and solves the problems that we pointed out earlier when discussing the counterfactual theory of causation. Instead of assessing counterfactuals like 1) George would not have plunged into the East River had he not jumped off the Brooklyn Bridge, and 2) George would not have jumped off the bridge had he not plunged into the East River, we should assess counterfactuals about manipulations: 1') George would not have plunged into the East River had he been independently manipulated to not jump off the Brooklyn Bridge, and 2') George would not have jumped off the bridge had he been independently manipulated not to have plunged into the East River. The difference is in how we interpret "independently manipulated." In the case of 2', we mean if we assign George to not plunging but leave everything else as it was, e.g., if we catch George just before he dunks. In this way of conceiving of the counterfactual, George *would* have jumped off the bridge, and so we can recover the asymmetry of causation once we augment the counterfactual theory with the idea of an independent manipulation, as Woodward argues.

4 Conclusion

Although the whirlwind tour in this short article is woefully inadequate, the references below (and especially their bibliographies) should be sufficient to point interested readers to the voluminous literature on causation produced in the last 30 years. Although vast and somewhat inchoate, it is safe to say that no reductive analysis of causation has emerged from this literature still afloat and basically sea-worthy. What I have described as the recent interdisciplinary theory of causation takes direct causation as a primitive, defines intervention from direct causation, and then connects causal systems to probabilities and statistical evidence through axioms, including the Causal Markov Axiom. Although it provides little comfort for those hoping to analyze causation Socratically, the theory does open the topic of causal epistemology in a way that has affected statistical and scientific practice, hopefully for the better. Surely that is some progress.

5 References

Bell, J. (1964). On the Einstein-Podolsky-Rosen Paradox." *Physics* 1: 195-200

--- (1966). On the Problem of Hidden Variables in Quantum Mechanics, *Reviews of Modern Physics*, 38:447-52.

Blalock, H. (1961). *Causal Inferences in Nonexperimental Research*. University of North Carolina Press, Chapel Hill, NC.

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

--. (1989). *Nature's Capacities and their Measurement*. Oxford, New York, Clarendon Press; Oxford University Press.

-- (2002). Against Modularity, the Causal Markov Condition, and Any Link Between the Two. *British Journal for Philosophy of Science*, 53: 411-453.

Dowe, P. (2000). Causality and Explanation. *British Journal for Philosophy of Science*, 51: 165-174.

Eells, E. (1991). *Probabilistic Causality*. Oxford University Press.

Glymour, C. (2004, forthcoming). Review of James Woodward, *Making Things Happen: A Theory of Causal Explanatio,* to appear in *British Journal for Philosophy of Science*,

Glymour, C., and Cooper, G. (1999). *Computation, Causation, and Discovery*. AAAI Press and MIT Press, Cambridge, MA.

Good, I.J. (1961,1962). A causal calculus I & II. *British Journal for Philosophy of Science*, 11: 305-18, and 12: 43-51.

Granger, C. (1969). Investigating Causal Relations By Econometric Models and Cross-Spectral Methods. *Econometrica* 37:424-438.

Hausman, D. (1984). Casual Priority, Nous, 18: 261-79

-- (1998). *Causal Asymmetries*. Cambridge University Press.

Halpern, J. and Pearl, J. (2002). Actual Causality, *IJCAI Proceedings*.

Hitchcock, C. (2001). The Intransitivity of Causation Revealed in Equations and Graphs,' *Journal of Philosophy*, 98: 273 - 299

Hitchcock, C. (2003). Of Humean Bondage, *British Journal for Philosophy of Science*, 54: 1-25.

Holland, P. (1986). Statistics and Causal Inference, *Journal of the American Statistical Association*, 81: 945-60.

Hoover, K. (2001) *Causality in Macroeconomics*, Cambridge University Press, New York.

Kiiveri, H. and T. Speed (1982) Structural analysis of multivariate data: a review. *Sociological Methodology*, ed. S. Leinhardt. Jossey-Bass.

Kiiveri, H., Speed, T., and Carlin, J. (1984). Recursive causal models. *Journal of the Australian Mathematical Society* **36**: 30-52.

Lauritzen, S. (1999). *Graphical Models*, Oxford University Press.

Lewis, D. (1973). *Counterfactuals*. Harvard University Press.

-- (1973). Causation, *Journal of Philosophy*, 70: 556-67.

-- (2000). Causation as Influence, *Journal of Philosophy*, 97: 182-197.

Mackie, J. (1974). *The Cement of the Universe*. Oxford University Press, New York.

McKim, S., and Turner, S. (1997). *Causality in Crisis? Statistical methods and the Search for Causal Knowledge in the Social Sciences*. University of Notre Dame Press.

Meek, C., and Glymour, C. (1994). Conditioning and Intervening. *British Journal for Philosophy of Science*, 45:1001-21.

Papineau, D. (1985). Causal Asymmetry. *British Journal for Philosophy of Science*, 36: 273 - 289.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems***.** Morgan and Kaufman, San Mateo

-- (2000). Causality: Models, Reasoning, and Inference. Cambridge University Press.

Reichenbach, H. (1956). *The Direction of Time*, University of California Press.

Rubin, D. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* 66, 688-701.

Salmon, W. (1984). *Scientific Explanation and the Causal Structure of the World*. Princeton University Press.

Simon, H. (1954).Spurious correlation: a causal interpretation. *JASA*. 49, 467-479.

Spirtes, P., Glymour, C., and Scheines, R. (2000). Causation, Prediction, and Search, 2nd Edition. MIT Press.

Spohn, W. (1983). Deterministic and probabilistic reasons and causes. *Erkenntnis*, 19:371-96.

Suppes, P. (1970). *A Probabilistic Theory of Causality*. North-Holland, Amsterdam.

Wold, H. (1954). Causality and Econometrics. *Econometrica*, 22: 162-77.

Woodward, J. (2003). *Making Things Happen: A Theory of Causal Explanation*. Oxford University Press.

Wright, S. (1934). The method of path coefficients. *Ann. Math. Stat.* 5, 161-215.

6 Acknowledgements

I thank Clark Glymour, Martha Harty, David Danks, and Mara Harrell for comments on early drafts and for many valuable discussions.

Sidebar 1: Event Causation vs. Causal Generalizations

Legal cases and accident investigations usually deal with a particular event and ask what caused it. For example, when in February 2003 the Space Shuttle Columbia burned up during re-entry, investigators looked for the cause of the disaster. In the end, they concluded that a chunk of foam insulation that broke off and hit the wing during launch was the cause of a rupture in the insulating tiles, which was the cause of the shuttle's demise during re-entry. Philosophers call this *event causation*, or *actual causation*, or *token-causation*.

Policy makers, statisticians and social scientists usually deal with *kinds* of events, like graduating from college, or becoming a smoker, or playing lots of violent video games. For example, epidemiologists in the 1950s and 1960s looked for the kind of event that was causing a large number of people to get lung cancer, and they identified smoking as a primary cause. Philosophers call this *type-causation*, or *causal generalization*, or *causation among variables*.

The properties of causal relationships are different for actual causation and for causal generalizations. Actual causation is typically considered transitive, anti-symmetric, and irreflexive. If we are willing to say that one event A, say the Titanic hitting an iceberg on April 12th, 1912, caused another event B, its hull ripping open below the water line and taking on water moments later, which in turn caused a third event C, it sinking a few hours later, then surely we should be willing to say that event A (hitting the iceberg) caused event C (sinking). So actual causation is transitive.[6] It is anti-symmetric because of how we view time. If a particular event A caused a later event B, then B did not cause A. Finally, single events don't cause themselves, so causation between particular events is irreflexive.

Causal generalizations, however, are usually but not always transitive, definitely not anti-symmetric and definitely not irreflexive. In some cases causal generalizations *are* symmetric, for example, confidence causes success, and success causes confidence, but in others they are not, for example, warm weather causes people to wear less clothing, but wearing less clothing doesn't cause the weather to warm. So causal generalizations are

[6] Plenty of philosophers disagree, for example see the work of Christopher Hitchcock.

asymmetric, not anti-symmetric, like actual causation. When they are symmetric, causal generalizations are reflexive. Success breeds more success, etc.

Sidebar 2: Overdetermination and Pre-emption

A spy, setting out to cross the desert with some key intelligence, fills his canteen with just enough water for the crossing and settles down for a quick nap. While he is asleep, Enemy A sneaks into his tent and pokes a very small hole in the canteen, and a short while later enemy B sneaks in, and adds a tasteless poison. The spy awakes, forges ahead into the desert, and when he goes to drink from his canteen discovers it is empty and dies of thirst before he can get water. What was the cause of the spy's death? According to the counterfactual theory, neither enemy's action caused the death. If enemy A hadn't poked a hole in the canteen, then the spy still would have died by poison. If enemy B hadn't put poison into the canteen, then he still would have died from thirst. Their actions overdetermined the spy's death, and the pinprick from enemy A pre-empted the poison from enemy B.

In the beginning of the movie Magnolia, a classic causal conundrum is dramatized. A 15-year-old boy goes up to the roof of his 10-story apartment building, ponders the abyss and jumps to his death. Did he commit suicide? It turns out that construction workers had installed netting the day before that would have saved him from the fall, but as he is falling past the 5th story, a gun is shot from inside the building by his mother and the bullet kills the boy instantly. Did his mother murder her son? As it turns out, his mother fired the family rifle at his drunk stepfather but missed and shot her son by mistake. She fired the gun every week at approximately that time after their horrific regular argument which the boy cited as his reason for attempting suicide, but the gun was never usually loaded. This week the boy secretly loaded the gun without telling his parents, presumably with the intent of causing the death of his stepfather. Did he, then, in fact commit suicide, albeit unintentionally?

Sidebar 3: The asymmetry of causation through causal connection

Two variables A and B are "causally connected" is either A is a cause of B, B a cause of A, or a third variable causes them both. If causation is transitive, then it turns out that everything causally connected to X is connected to its effects, but not everything connected to Y is connected to its causes. When X → Y, everything causally connected to X is causally connected to Y (Figure 7 -A), but something causally connected to Y is *not* necessarily causally connected to X (Figure 7-B).

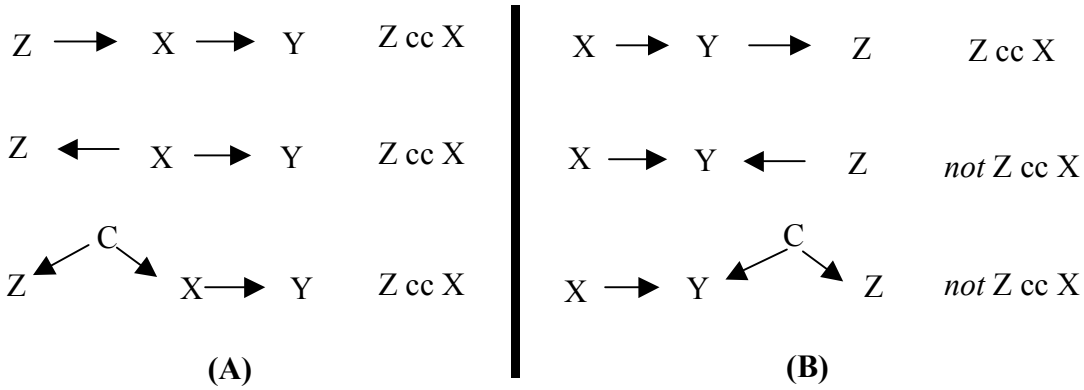

Figure 7: The Asymmetry in the Transitivity of Causal Connection