# Section 7: Causal and Counterfactual Reasoning

## *Chapter 29: Causal Thinking*

LANCE J. RIPS[1]

One damn thing leads to another. I forget to open the garage door this morning, back my car into the door, and splinter it. The actions we perform cause other events – my backing up causes the splintering. But events of other kinds – nonactions – have their effects, too. With no help from me, last night's storm caused a branch to fall from a tree, putting a hole in my roof.

Much as we might like to forget them, we often keep track of events like these and the causes that unite them. Although we might not have predicted these events, we can remember and reconstruct part of the causal sequences after they occur. In retelling the events of last summer, for example, we tend to relate the events in forward causal order, starting, say, at the beginning of our trip to Virginia in May and proceeding chronologically. If we want to mention other kinds of events from the same period, such as our summer work experiences, we may start again at the beginning of the summer, moving along the events in a parallel causal stream (Barsalou 1988). We also remember fictional stories in terms of the causal changes that compose their main plot line, remembering less about events falling on deadend side plots (Trabasso and Sperry 1985). We sometimes attribute causal powers to concrete objects as well as to events, but we can understand this sort of talk as an abbreviation for event causation. If Fred caused the glass to break that's because one of Fred's actions – maybe his dropping it – caused the breaking. I'll take event causation as basic in this article on the strength of such paraphrases.

We remember causes and effects for event types as well as for event tokens. Ramming heavy objects into more fragile ones typically causes the fragile items damage; repeating phone numbers four or five times typically causes us to remember them for awhile. Negotiating routine events (e.g., Schank and Abelson 1977), constructing explanations (e.g., Lewis 1986), and making predictions all require memory for causal relations among event categories. Causal generalities underlie our concepts of natural kinds, like daisies and diamonds (e.g., Ahn and Kim 2000; Barton and Komatsu 1989; Gelman and Wellman 1991; Keil 1989; Rehder and Hastie 2001; Rips 1989, 2001) and support our concepts of artifacts like pianos or prisms. Our knowledge of how beliefs and desires cause actions in other people props up our own social activities (e.g., Wellman 1990).

The importance of causality is no news. Neither are the psychological facts that we attribute causes to events, remember the causes later, and reason about them – although, as usual, controversy surrounds the details of these mental activities. Recently, though, psychologists seem to be converging on a framework for causal knowledge, prompted by earlier work in computer science and philosophy. Rhetorical pressure seems to be rising to new levels among cognitive psychologists working in this area: For example, "until recently no one has been able to frame the problem [of causality]; the discussion of causality was largely based on a framework developed in the eighteenth century. But that's changed. Great new ideas about how to represent causal systems and how to learn and reason about them have been developed by philosophers, statisticians, and computer scientists" (Sloman 2005: vii). And at a psychological level, "we argue that these kinds of representations [of children's knowledge of causal structure] and learning mechanisms can be perspicuously understood in terms of the normative mathematical formalism of directed graphical causal models, more commonly known as Bayes nets.... This formalism provides a natural way of representing causal structure, and it provides powerful tools for accurate

prediction and effective intervention" (Gopnik et al. 2004: 4).

It's a little unfair to catch these authors in mid rhetorical flight. But the claims for these formalisms do provoke questions about how far they take us beyond the simple conclusions I've already mentioned. Kids and adults learn, remember, and apply causal facts. As a card-carrying CP (i.e., cognitive psychology) member, I believe that kids and adults therefore mentally represent these facts. But what's new here that further illuminates cognitive theorizing? Here's the gloomy picture: The new methods are at heart data-analytic procedures for summarizing or approximating a bunch of correlations. In this respect, they're a bit like factor analysis and a whole lot like structural equation modeling. (If you think it surprising that psychologists should seize on a statistical procedure as a model for ordinary causal thinking, consider that another prominent theory in this area is Kelley's [1967] ANOVA model; see the section on *Causation from Correlation*, and Gigerenzer 1991.) The idea that people use these methods to induce and represent causality flies in the face of evidence suggesting that people aren't much good at normatively correct statistical computations of this sort (e.g., Tversky and Kahneman 1980). Offhand, it's much more likely that what people have are fragmentary and error-prone representations of what causes what.

The rosier picture is the one about "great new ideas."

The jury is still out, and I won't be resolving this issue here. But sorting out the claims for the new causal representations highlights some important questions about the nature of causal thinking.

## How Are Causal Relations Given to Us?

Here's a sketch of how a CD player works (according to Macaulay 1988): A motor rotates a spindle that rotates the CD. As the CD turns, a laser sends a beam of light through a set of mirrors and lenses onto the CD's surface. The light beam lands on a track composed of reflecting and nonreflecting segments that have been burned onto the CD. The reflecting segments bounce the light beam back to a photodiode that registers a digital "on" signal; the nonreflecting segments don't bounce the light back and represent an "off" signal. The pattern of digital signals is then converted into a stereo electrical signal for playback.

You could remember this information in something like the form I just gave you – an unexciting little narrative about CD players. But the new psychological approach to causal knowledge favors directed graphs like Figure 1 as mental representations – "causal maps" of the environment (Gopnik et al. 2004). This graph contains nodes that stand for event types (e.g., the CD player's motor rotating or not rotating, the CD turning or not turning) and directed links that stand for causal connections between these events (the motor rotating causes the CD's turning; the laser producing a beam and the mirror-lens assembly focusing the beam jointly cause the beam to hit the CD's surface). Of course, no one disputes the fact that people can remember some of the information these diagrams embody. Although people can be overconfident about their knowledge of mechanical devices like this one (Rozenblit and Keil 2002), they're nevertheless capable of learning, say, that the CD player's motor causes the CD to turn. What's not so clear is how they acquire this cause-effect information, how they put the component facts together, and how they make inferences from such facts. In this section, we'll consider the acquisition problem, deferring issues of representation and inference till the second part of this chapter.

### *Causation in Perception*

You're not likely to get much of the information in Figure 1 by passively observing a CD player, unless you already know about the nature of similar devices. But sometimes you do get an impression of cause from seeing objects move. Repeated sightings of an event of type $E_1$ followed by an event of type $E_2$ may provide evidence that $E_1$ causes $E_2$. Rather weak evidence, but evidence nonetheless. When we later see an example of the same sequence, we can infer the causal link. But psychologists sometimes claim there is a more intimate perception of cause in which an observer directly experiences one event causing another.

#### PERCEPTUAL STUDIES

In a famous series of demonstrations, Michotte (1963) rigged a display in which a square appeared to move toward a second square and to stop abruptly when they touched. If the second square then began to move within a fixed

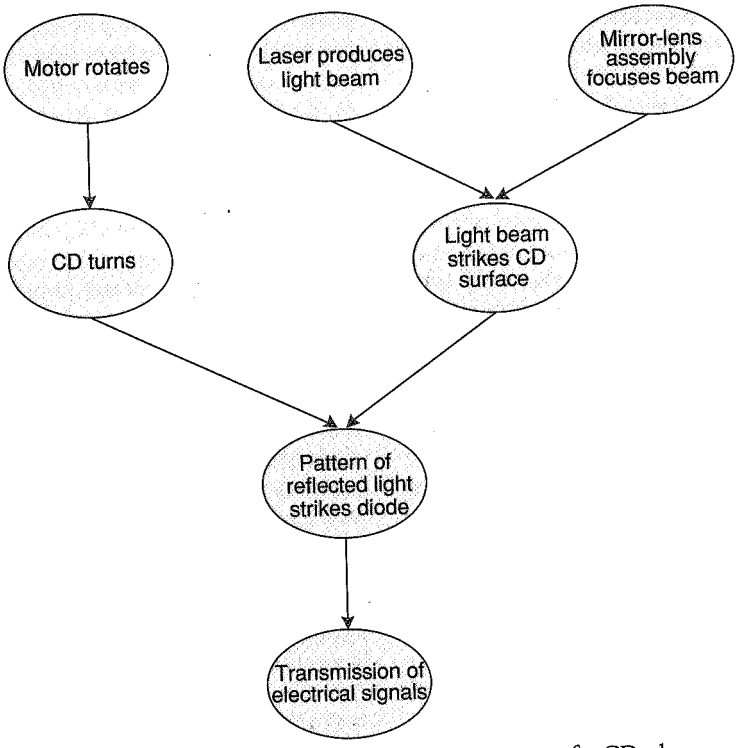


Figure 1. A directed graph representing the operation of a CD player, based on text by Macaulay (1988).

interval of the touching and at a speed similar to that of the first square, observers reported the first square causing the second to move, *launching* it.

Michotte's extensive experiments aimed to isolate the purely perceptual conditions that produce this immediate impression of causality, but there's a paradoxical quality to his efforts. The first square in the display doesn't actually cause the second to move. The displays showed 2-D projections of simple geometrical forms whose movements could be carefully controlled behind the scenes. (In those days before lab computers, Michotte engagingly created his displays using striped disks rotating behind slits or using pairs of moving slide projectors.) The goal was therefore not to determine when people correctly detect causal relations in their environment but instead to uncover the cues that lead them to report causality.[2] Michotte himself discusses a number of situations in which people report one event causing another, even though the interaction is physically unlikely or impossible. In one such case, a square A moves at 30 cm/s and comes into contact with another square B, which is already moving at 15 cm/s. If A comes to a halt and B moves off at a *slower* pace than before (7.5 cm/s), observers report a causal effect. "Such cases are particularly interesting in that they show that causal impressions arise as soon as the psychological conditions of structural organization are fulfilled, and indeed that they can arise even in situations where we know from past experience that a causal impression is a downright impossibility" (Michotte 1963: 71). Michotte's project attempted to explain these causal impressions in noncausal terms: His descriptions of the crucial stimulus conditions don't presuppose one object causally influencing another. He believed that people's impression of causality arises as their perceptual systems try to resolve a conflict (e.g., in the launching event) between the initial view of the first square moving and the second square stationary and the final view of the first square stationary and the second moving. The resolution is to see the movement of the first object extending to the second, which Michotte called "ampliation of the movement" (which, I hope, sounds better in French).

Michotte (1963: 351–352) believed that this resolution "enables us to understand why, when

such a structure is established, participants can communicate adequately what they perceive only by saying that they *see* the [initially moving] object make the second go forward." Why? The obvious answer would be that this perceptual situation is one that real objects produce when they undergo causal interactions. The resolution that takes place in the experimental displays reminds the observers, perhaps unconsciously, of what happens when they view causal comings and goings in the ordinary environment, and they therefore interpret it the same way. But this answer is one Michotte rejects, since he consistently denies that the launching effect is due to acquired knowledge. This is why physically impossible cases, like the one described in the previous paragraph, are important to him: They seem to rule out the possibility that observers are making an inference to causality based on experience.

The easiest way to understand Michotte's theory (though not in terms he used) is as the claim that people have a built-in causality detector that is triggered by the conditions he attempted to describe. Since the detector is presumably innate, its operations don't depend on learning from previous experience. Moreover, the detector responds reliably but not perfectly. Toads dart at insects in their visual fields but can be tricked into darting at moving black-on-white or white-on-black spots, according to the old ethology chestnut (e.g., Ewert 1974). In the same way, whenever the movement of an object "extends" to a second, people receive the impression of causality, whether or not the first object actually causes the second to move.

But this approach, like some moving spots, is hard to swallow. Although Michotte stressed that observers spontaneously report the events in causal language – for example, that "the first square pushed the second" – the impression of causality doesn't seem as immediate or automatic as typical perceptual illusions. We can't help but see the apparent difference in line length in the Muller-Lyer illusion or the apparently bent lines in the Poggendorf and Hering figures (see, e.g., Gregory 1978, for illustrations of these). And toads, so far as we know, can't help unleash their tongues at moving specks. But Michotte's demonstrations allow more interpretative leeway.

Suppose Michotte was right that people possess an innate detector of some sort that's broad enough to be triggered by the displays his participants report as causal. The detector, of course, produces false positive responses to some displays that are in fact noncausal (e.g., Michotte's displays), and it produces false negative responses to some causal ones (e.g., reflections of electromagnetic rays in an invisible part of the spectrum). So what the detector detects is not (all or only) causal interactions but perhaps something more like abrupt transitions or discontinuities in the speed of two visible objects at the point at which they meet. This would include both the normal launching cases and the causally unlikely or impossible ones, such as slowing on impact. Nor do we ultimately take the output of the detector as indicating the presence of a causal interaction. In the case of Michotte's demos, for example, we conclude that no real causal interaction takes place between the squares, at least when we become aware of what's going on behind the smoke and mirrors. The issue of whether we *see* causality in the displays, then, is whether there's an intermediate stage between the detector and our ultimate judgment, a stage that is both relevantly perceptual and also carries with it a causal verdict. Because these two requirements pull in opposite directions, the claim that we can see causality is unstable.

Here's an analogy that may help highlight the issue. People viewing a cartoon car, like the ones in the Disney film *Cars*, immediately "see" the cartoon as a car (and report it as a car), despite the fact that it is physically impossible for cars to talk, to possess eyes and mouths, and to move in the flexible way that cartoon cars do. Although I don't recommend it, you could probably spend your career pinning down the parameter space (e.g., length-to-width ratios) within which this impression of carness occurs. But there isn't enough evolutionary time since the invention of cars in the 19th century for us to have evolved innate car detectors. The fact that we immediately recognize cartoons as cars even when they possess physically impossible properties can't be evidence for innate car perception. Michotte's evidence seems no stronger as support for innate cause detection. Although it's an empirical issue, I'm willing to bet that the impression of carness generated by the cartoon cars is at least as robust as the impression of causality generated by launching displays.

Causality is an inherently abstract relation – one that holds not only between moving physical objects but also between subatomic particles, galaxies, and lots in the middle – and this abstractness makes it difficult to come up with a plausible theory that would have us perceiving it directly, as opposed to inferring it from

more concrete perceived information.[3] There's no clear way to defeat the idea that "when we consider these objects with the utmost attention, we find only that the one body approaches the other, and the motion of it precedes that of the other without any sensible interval" (Hume 1739/1967: 77).

### DISSOCIATION BETWEEN PERCEIVED AND INFERRED CAUSALITY

More recent evidence suggests that people's judgments about perceived causality are independent of some of the inferences they make about cause. Investigators have taken these dissociations to suggest that Michotte (1963) was right that perceived causality is an innate module. One such study (Roser, Fugelsang, Dunbar, Corballis, and Gazzaniga 2005) employed two split-brain patients, presenting causal tasks to the patients' right or left hemispheres. In one task, the patients saw Michotte-type launching events that varied in the spatial gap between the two objects at the moment the second object began to move and, also, the time-delay between the point at which the first object stopped and the second object began moving. Both spatial gaps and time delays tend to weaken the impression of perceived causality in normal participants. And so they did in the split-brain patients, but with an important qualification. The patients had to choose whether the first object appeared to cause the second to move or whether the second object moved on its own, and their positive "cause" judgments were more frequent when there was no delay and no gap. This difference appeared, however, only when the patients' right hemisphere processed the display. Left-hemisphere processing showed no difference between conditions. A second task asked the same split-brain patients to solve a problem in which they had to use the statistical co-occurrence between visually presented events to decide which of two switches caused a light to come on. Patients were more often correct in this task when the displays presented the information to their left hemispheres than when they presented it to their right hemispheres.

Split-brain patients may process causal information in atypical ways, but investigators have found similar dissociations with normal participants. Schlottmann and Shanks (1992, Experiment 2) varied the temporal gap within launching events (as in Roser et al. 2005) and also the contingency that existed across trials between whether the first object moved and whether the second object moved. On some series of trials, the first object's moving was necessary and sufficient for the second object to move; on others, the second object could move independently of the first. Participants made two types of judgments on separate trials within these series: how convincing a particular collision appeared and whether the collisions were necessary for the second object to move. Schlottmann and Shanks found an effect of delay but no effect of contingency on judgments of the display's convincingness. Judgments of necessity, however, showed a big effect of contingency and a much smaller effect of delay.

These dissociations suggest – what should become clear in the course of this chapter – that causal thinking is not of one piece. Some causal judgments depend vitally on detailed perceptual processing, while others depend more heavily on schemas, rules, probabilities, and other higher-order factors. What's not so clear is whether the dissociations also clinch the case for a perceptual causality detector. The right hemispheres of Roser et al.'s (2005) split-brain patients could assess the quality of launching events even though they were unable to evaluate the impact of statistical independencies. But this leaves a lot of room for the influence of other sorts of inference or association on judgments about launching. Suppose, for example, that launching judgments depend on whether observers are reminded of real-world interactions of similar objects. Unless the right hemisphere is unable to process these reminders, inference could still influence decisions about launchings. Similarly, Schlottmann and Shanks's (1992) finding shows that observers can ignore long-run probabilities in assessing the convincingness of a particular collision, but not that they ignore prior knowledge of analogous physical interactions.

### STUDIES OF INFANTS

Developmental studies might also yield evidence relevant to Michotte's claim, since if the ability to recognize cause is innate, we should find infants able to discriminate causal from non-causal situations. The evidence here suggests that by about six or seven months, infants are surprised by events that violate certain causal regularities (Kotovsky and Baillargeon 2000; Leslie 1984; Leslie and Keeble 1987; Oakes 1994). In one such study, for example, Kotovsky and Baillargeon first showed seven-month-olds static displays containing a cylinder and a toy bug, either with a thin barrier separating them (no-contact

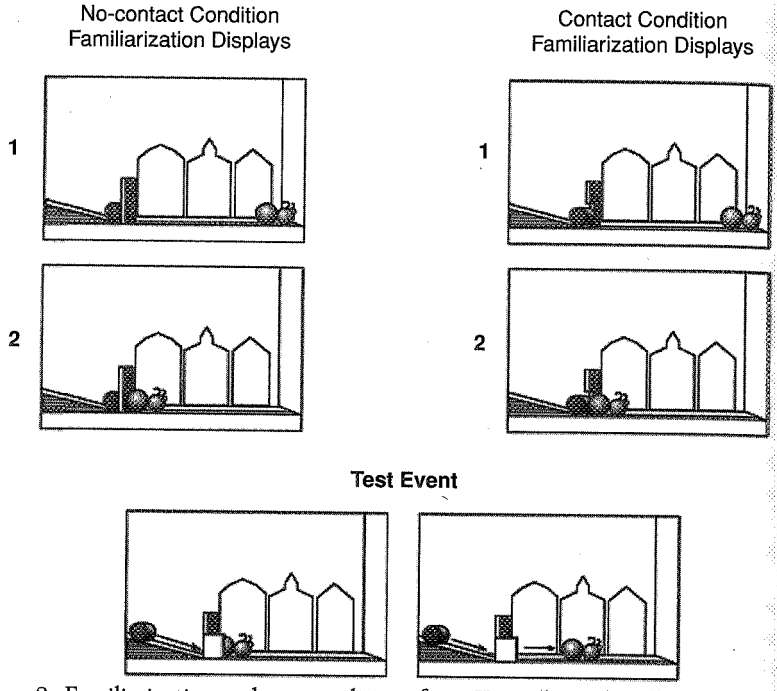


Figure 2. Familiarization and test conditions from Kotovsky and Baillargeon (2000).

condition) or with a partial barrier that did not separate them (contact condition). Figure 2 displays these two conditions at the left and right, respectively. A screen then hid the position that contained the barrier or partial barrier. In the experiment's test phase, the infants saw the cylinder roll down a ramp and go behind the screen, as shown at the bottom of Figure 2. The screen hid what would be the point of impact, but if the bug moved as if the cylinder had struck it, the infants looked longer in the no-contact than in the contact condition. If the bug failed to move, infants showed the opposite pattern of looking.

At seven months,[4] then, infants appear to discriminate some cases in which simple launching events will and won't occur, but should we take this as evidence for innate perception of causality? Unfortunately, there seems to be no evidence that would allows us to compare directly the class of interactions that Michotte's participants report as causal with the class that infants react to. It would be useful to know, in particular, whether the "impossible" displays that Michotte's observers report as causal are also ones to which infants give special attention. What we do know, however, is that infants take longer than seven months to recognize causal interactions even slightly more complex than simple launching. For example, at seven months they fail to understand situations in which one object causes another to move in a path other than dead ahead, situations that adults report as causal (Oakes, 1994).

If the classes of interactions that adults and infants perceive as causal are not coextensive, this weakens the evidence for innate, modular perception of causality. You could maintain that the perceptual impression of causality changes with experience from an innate starting point of very simple causal percepts, such as dead on launchings, but this opens the door to objections to the very idea of directly perceiving cause. If learning can influence what we see as a causal interaction, then it seems likely that top-down factors – beliefs and expectations – can affect these impressions. Perhaps the learning in question is extremely local and low level. But if not – if observers' impressions of cause change because of general learning mechanisms – then this suggests that the impressions are a matter of inference rather than direct perception. Much the same can be said about evidence that seven-month-olds' reaction to launching events depends on whether the objects are animate or inanimate (Kotovsky and Baillargeon 2000). The animacy distinction presumably depends on higher-level factors, not just on the spatiotemporal parameters Michotte isolated (see Saxe and Carey 2006 for a review).

Of course, uncertainty about the evidence for direct perception of causality needn't affect the claim that the concept of causality is innate (see the section on causal primitives later in this chapter). Children may have such a concept but be initially unsure exactly what sorts of perceptual data provide evidence it applies. Moreover, nonperceptual, as well as perceptual, data may trigger such a concept; in fact, most theories of causality in psychology have avoided tying cause to specifically perceptual information. These theories take seriously the other aspect of Hume's (1739/1967) view, trying to account for judgments of causality in terms of our experience of the co-occurrence of events. Does recent research shed any light on this possibility?

## *Causation from Correlation*

Even if we can literally perceive causality in some situations, we have to resort to indirect methods in others. A careful look at a CD player's innards can't disclose the causal link between the reflected pattern of light and the transmission of sound signals at the bottom of Figure 1. We may see the reflected light and hear the resulting sound, but we don't have perceptual access to the connection between them. Similarly, we can't see atmospheric pressure influencing the boiling point of a liquid or a virus producing a flu symptom or other people's beliefs motivating their actions. Experiments in science would be unnecessary if all we had to do to isolate a causal mechanism is look.

Scientists, of course, aren't the only ones in need of hidden causal facts. We need to predict how others will behave if we want to enlist them in moving a sofa. We need to know what buttons to press if we want to make a cell phone call or record an opera broadcast or adjust the drying cycle to keep from scorching our socks. We need to know which foods are likely to trigger our allergy, which windows are best for which plants, which greetings will produce another greeting versus a stunned silence or a slap in the face. We can sometimes rely on experts to tell us about the hidden causes. Allergists are often good on allergies, botanists on plants, and Miss Manners on manners. But sometimes we have to proceed on our own, and the question is how ordinary people cope with the task of recognizing causal relationships when they can't look them up. The answer that psychologists have usually given to this question is that people operate from bottom up, observing the temporal co-occurrence of events and making an inductive inference to a causal connection. They might passively register the presence or absence of a potential cause and its effects or they may actively intervene, pressing some buttons to see what happens. In either case, they decide whether a cause-effect link is present on the basis of these results. This section considers the more passive route to discovering causes, and the next section looks at the more active one.

### CAUSE, CONTRAST, CORRELATION

If we suspect event type C causes event type E, we should expect to find E present when C is present and E absent when C is absent. This correlation might not be inevitable even if C really is a cause of E. Perhaps E has an alternative cause C′; so E could appear without C. Or perhaps C is only a contributing cause, requiring C″ in order to produce E; then C could appear without E. But if we can sidestep these possibilities or are willing to define *cause* in a way that eliminates them, then a correlation between C and E may provide evidence of a causal relation. Codifying this idea, Mill (1874) proposed a series of well-known rules or canons for isolating the cause (or effect) of a phenomenon. The best known of these canons are the method of agreement and the method of difference. Suppose you're looking for the cause of event type E. To proceed by the method of agreement, you should find a set of situations in which E occurs. If cause C also occurs in all these situations but no other potential cause does, then C causes E. To use the method of difference, which Mill regarded as more definitive, you should find two situations that hold constant all but one potential cause, C, of E. If E is present when C is present, and E is absent when C is absent, then C causes E.

Psychologists have mostly followed Mill's canons in their textbooks and courses on scientific methods.[5] If you're a victim of one of those courses, you won't find it surprising that psychological theories of how nonscientists go about determining cause-effect relations reflect the same notions:

> The inference as to where to locate the dispositional properties responsible for the effect is made by interpreting the raw data . . . in the context of subsidiary information from experiment-like variations of conditions. A näive version of J. S. Mills' method of difference provides the basic analytic tool. The effect is attributed to that condition which is present when the effect is present and which

**Table 1: Two Contrasts for Assessing the Presence of a Causal Relation**

a.

| | *This Occasion* | | *Other Occasions* | |
|---|---|---|---|---|
| | *Calvin* | *Other People* | *Calvin* | *Other People* |
| tango | 1 | 0 | 1 | 0 |
| other dances | 1 | 0 | 1 | 0 |

b.

| | *This Occasion* | | *Other Occasions* | |
|---|---|---|---|---|
| | *Calvin* | *Other People* | *Calvin* | *Other People* |
| tango | 1 | 1 | 1 | 1 |
| other dances | 0 | 0 | 0 | 0 |

1's indicate that a person likes a particular dance on a given occasion; 0's indicate not liking to dance.

> is absent when the effect is absent. (Kelley 1967: 194)

As an example (similar to one from Cheng and Novick 1990), suppose you know that Calvin danced the tango last Thursday. To find out the cause of this event, you need to examine potential causes that the outcome suggests: Maybe it was a disposition of Calvin's, maybe it was the tango, maybe it was something about this particular occasion. To figure out which of these potential causes was at work, you mentally design a study in which the three causes are factors. The design will look something like what's in Table 1. The 1's in the cells stand for somebody dancing on a particular occasion, and the 0's stand not dancing. If the pattern of data looks like what's in Table 1a, we have an effect for the person but no effects for either the occasion or the type of dance; so we might conclude that the reason Calvin danced the tango on this occasion is that he just likes dancing. By contrast, if the data come out in the form of Table 1b, where Calvin and others don't do other kinds of dancing, but everyone dances the tango, we might conclude that it was the tango that caused Calvin's dancing.

Kelley's (1967) ANOVA (analysis of variance) theory aimed to explain how individuals determine whether their reaction to an external object is due to the object itself (e.g., the tango) or to their own subjective response, and the theory focused on people, objects, times, and "modalities" (different ways of interacting with the entity) as potential factors. Cheng and Novick (1990, 1992) advocated a somewhat more flexible approach in which people choose to consider a set of potential factors on pragmatic grounds: "Contrasts are assumed to be computed for attended dimensions that are present in the event to be explained" (1990: 551). According to this theory, people also determine causation relative to a particular sample of situations, a "focal set," rather than to a universal set. Within these situations, people calculate causal effectiveness in terms of the difference between the probability of the effect when the potential cause is present and the probability of the effect when the same potential cause is absent:

$$(1) \quad \Delta P = \text{Prob}(\text{effect} \mid \text{factor}) - \text{Prob}(\text{effect} \mid \sim\text{factor}),$$

where *Prob*(effect | factor) is the conditional probability of the effect given the presence of the potential causal factor and *Prob*(effect | ~factor) is the conditional probability of the effect given the absence of the same factor. When this difference, $\Delta P$, is positive, the factor is a contributory cause of the effect; when it's negative, the factor is an inhibitory cause; and when it's zero, the factor is not a cause. Cheng and Novick also distinguish causes (contributory or inhibitory) from "enabling conditions" – factors whose $\Delta P$ is undefined within the focal set of situations (because they are constantly present or constantly absent) but that have nonzero $\Delta P$ in some other focal set.

We can illustrate some of these distinctions in the Table 1 results. In Table 1a, $\Delta P = 1$ for Calvin versus other people, but 0 for the object and occasions factors. So something about

Calvin is a contributory cause of his dancing the tango at that time, and the tango and the occasion are noncauses. In the Table 1b data, the object (dance) factor has a $\Delta P$ of 1, whereas the person and occasion factors have $\Delta P$'s of 0; so the tango causes the event. Reversing the 0's and 1's in Table 1b, so that Calvin and others never dance the tango but always dance other dances, will produce a $\Delta P$ of $-1$. In this case, the tango is an inhibitory cause. A factor – perhaps, music – that is present in all the situations in the focal set considered here would be an enabling condition if it turned out to have a positive $\Delta P$ in a larger sample of situations in which it was present in some and absent in others. The results in Table 1 are all-or-none, but the $\Delta P$ measure obviously generalizes to situations in which the effect can occur within each cell sometimes but not always.

Related notions about cause derive from work on associative learning. Creatures learning that, say, a shock often follows a tone are remembering contingency information about the tone and shock (or the pain or fear that the shock creates – sorry, animal lovers, but these aren't my experiments). A number of researchers have proposed that this primitive form of association might provide the basis for humans' causal judgments (e.g., Shanks and Dickinson 1987; Wasserman, Kao, Van Hamme, Katagiri, and Young 1996). Data and models for such learning suggest that this process may be more complex than a simple calculation of $\Delta P$ over all trials. In particular, the associative strength between a specific cue (e.g., tone) and an unconditioned stimulus (shock) depends on the associative strength of other cues (lights, shapes, colors, etc.) that happen to be in play. The associative strength for a particular cue is smaller, for example, if the environment already contains stronger cues for the same effect. If these associative theories are correct models for judgments about a specific potential cause, then such judgments should depend on interactions with other potential causes, not just on "main effect" differences like those of the ANOVA model or $\Delta P$. Evidence for these interactions in causal judgments appears in a number of studies (e.g., Chapman and Robbins 1990; Shanks and Dickinson 1987).[6] However, $\Delta P$-based theories can handle some of these results if participants compute $\Delta P$ while holding other confounded factors constant (a conditional $\Delta P$, see Cheng 1997; Spellman 1996). Also, under certain conditions (e.g., only one potential cause present), associative theories sometime reduce to $\Delta P$ (Chapman and Robbins 1990; Cheng 1997).[7] Because both associative and statistical models make use of the same bottom-up frequency information, we consider them together here (see the section on Power for more on interactions).

### LOTS OF CORRELATIONS

The same textbooks on methodology that extol Mill's canons of causal inference also insist that a correlation between two variables can't prove that one causes the other. Because Mill's methods, the ANOVA theory, $\Delta P$, associative theories, and their variants all work along correlational lines, how can they provide convincing evidence for causation?[8] If these methods yield a positive result, there's always the possibility that some unknown factor confounds the relation between the identified cause and its effect. Maybe Calvin's love of dancing didn't cause his dancing the tango Thursday, but instead the cause was his girlfriend's insistence that he dance every dance on every occasion (in the Table 1a example). If these methods yield a negative result for some putative cause, there's always the possibility that some unknown factor is suppressing the first. The tango's special allure might surface if Calvin and his girlfriend hadn't crowded other couples off the dance floor. If we can't identify a cause (due to possible confounding) and we can't eliminate a potential cause (because of possible suppression), how can we make any progress with these correlational methods? Of course, the ANOVA theory and the $\Delta P$ theory (unlike Mill's methods) are intended as models of ordinary people's causal reckoning, and ordinary people may not consider confoundings or suppressors. Superstitious behavior may attest to their unconcern about spurious causes and noncauses, as might the need for the textbook warnings about these weak inferences. Even children, however, can reject confoundings under favorable conditions (Gopnik et al. 2004; Koslowski 1996: Ch. 6). So we seem to need an explanation for how people can go beyond correlation in their search for causes.

Although a single contrast or correlation between factors may not be convincing evidence, multiple correlations may sometimes reveal more about the causal set up. To see why this is so, let's go back to the CD diagram in Figure 1. Both the rotating motor and the laser beam influence the final transmission of electrical signals. So we would expect both the rotation of the motor and the presence of the laser beam to be correlated with the transmission. The correlation between the motor and the light beam, however, should be zero, provided no further

factors outside the diagram influence both of them. (If there is a power switch, for example, that controls both the motor and the laser, then, of course, there will be such a correlation. So imagine there are separate controls for present purposes.) Similarly, the diagram predicts that if we can hold constant the state of some of the variables in Figure 1, the correlation among other variables should go to zero. For instance, although there should be a correlation between whether the CD is rotating and transmission of signals, we should be able to break the correlation by observing only those situations in which the intermediate variable, the light striking the diode is constant For instance, when light is not striking the diode, there should be no correlation between the rotating and the transmission. The causal relations among the different parts of the diagram put restrictions on what is correlated with what. Working backward from the pattern of correlations, then, we may be able to discern which causal relations are consistent with these correlations and which are not. For example, the presence of a correlation between the rotation and the light beam would be a reason to think that the causal arrows in Figure 1 are incorrect. Statistical techniques like path analysis and structural equation modeling exploit systems of correlations in this way to test theories about the causal connections (e.g., Asher 1983; Klem 1995; Loehlin 1992).

There are limits to these methods, however, that are similar to those we noted in connection with single correlations (Cliff 1983). In the first place, there may still be confounding causes that are not among the factors considered in the analysis. In the setup of Figure 1, for example, we should observe a correlation between the light striking the diode and the transmission of signals, but there is no guarantee, based on correlations alone, that this is due to the direct effect of the diode on the signals (as the figure suggests). Rather, the correlation could be due to the effect of some third, confounding variable on both the diode and the signal. The same is obviously true for the rest of the direct connections that appear in the graph. Each direct connection is subject to exactly the same uncertainty about confoundings that we faced with single correlations. Second, the pattern of correlations can drastically underdetermine the causal structure. Consider, for example, a completely arbitrary set of correlations among four variables A, B, C, and D. The causal connections in Figure 3a (i.e., A has a direct causal effect on B, C, and D; B has a direct effect on C and D; and C has a direct effect on D) will be perfectly consistent with those correlations, whatever they happen to be. For example, a path analysis based on these connections will *exactly* predict the arbitrary correlations. Moreover, so will any of the other twenty-three models in which the position of the variables in this structure is permuted—for instance, the one in Figure 3b in which D directly causes C, B, and A; C directly causes B and A; and B directly causes A. These are *fully recursive* models in path-analysis terminology, and they always fit the data perfectly. Additional information beyond the correlations would be necessary to discriminate among these sets of possible causal connections (Klem 1995; see also Pearl 2000 for a discussion of Markov equivalent causal structures).

### CAUSAL MECHANISMS AND SCHEMAS

To compound these difficulties for the bottom-up, correlation-to-causation approach, the causal environment typically contains an enormous number of factors that could produce a given effect. Calvin, the tango, or the occasion may produce events that cause his dancing the tango on Thursday, but these factors are cover terms that contain many different potential causes: They serve as causal superordinate categories. Not all of Calvin's dispositions would plausibly cause him to dance, but this still leaves a seemingly unlimited number to choose from. Is the cause his showmanship, his athleticism, his musical talents, his religious fervor, his distaste of being a wallflower, his fear of letting down his girlfriend, . . . ? Moreover, we needn't stop at people, objects, and occasions, as we've already noted. Maybe it's his girlfriend's demands, maybe it's bribery by the DJ, maybe it's cosmic rays, maybe it's his therapist's hypnotic suggestion, maybe it's a disease (like St. Vitus dance), and so on. Since there is no end to the possibilities, there is no way to determine for each of them whether it is the cause, making a purely bottom-up approach completely hopeless.

We should again distinguish the plight of the scientist from the task of describing laypeople's causal search. Laypeople may take into account only a handful of potential causes and test each for a correlation with the effect. Although such a procedure might not be normatively correct, it may nevertheless be the recipe people follow in everyday life. But even if people use correlations over a restricted set of factors, an explanation of their causal reasoning would then also have to include an account at how they arrive at the restricted set. The factors they test are the factors

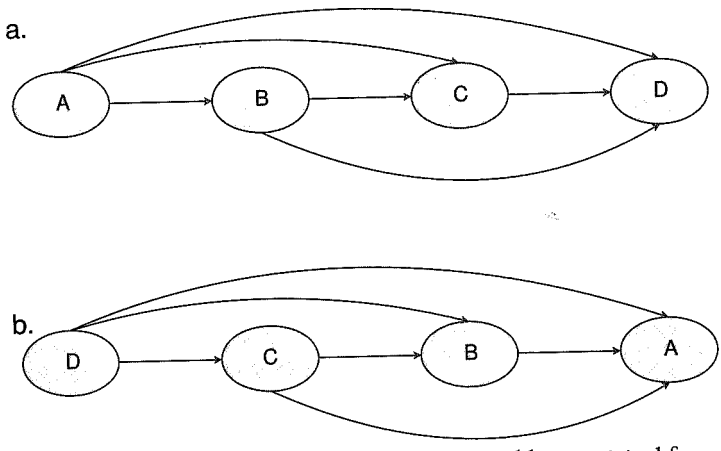


Figure 3. A hypothetical causal model for four variables in original form (a) and permuted form (b).

they attend to, of course, but what determines what they attend to? People's causal thinking often aims at explaining some phenomenon, where what needs explaining may be a function of what seems unusual or abnormal within a specific context (Einhorn and Hogarth 1986; Hilton 1988; Kahneman and Miller 1986). The explanation process itself depends on broadly pragmatic factors, such as the explainers' interest or point of view, the contrast class of explanations they have in mind, the intended audience for the explanation, and the availability of evidence, among others (Brem and Rips 2000; Hilton 1990; Lewis 1986; van Fraassen 1980). The same goes for determining "the cause" of a phenomenon, which is a disguised way of asking for the main cause or most important cause.

Evidence supports the notion that people's search for causes relies on information other than correlation. Ahn, Kalish, Medin, and Gelman (1995) asked participants what kinds of evidence they needed to determine the cause of an event like the one about Calvin. (Ahn et al. used some of the stimulus materials from Cheng and Novick 1990.) For example, participants had to write down questions that they would like to have answered in order to figure out the cause of (in this case) Calvin's *not* dancing the tango on this occasion. Ahn et al. predicted that if the participants were following the ANOVA or Cheng-Novick $\Delta P$ theory, they should be seeking information that fills out the rest of the design matrix in Table 1 – the kind of information they could use to compute experimental contrasts or $\Delta P$. Did other people dance the tango? Did Calvin dance other kinds of dances? Did Calvin dance the tango on other occasions? And so forth. What Ahn et al. found, though, is that participants asked these sorts of questions only about 10 percent of the time. Instead, participants asked what Ahn et al. call "hypothesis-testing" questions, which were about specific explanatory factors not explicitly mentioned in the description of the event. Participants asked whether Calvin had a sore foot or whether he ever learned the tango, and about similar sorts of common-sense causal factors. These hypothesis-testing questions showed up on approximately 65 percent of trials. Ahn et al. concluded that when people try to explain an event, they look for some sort of mechanism or process that could plausibly cause it. They have a set of these potential mechanisms available in memory, and they trot them out when they're trying to discover a cause.

People may also infer correlational information from their causal beliefs rather than the other way round. Psychologists have known since Chapman and Chapman's (1967; Chapman 1967) initial work on illusory correlations that causal expectancies can affect estimates of correlations (for reviews, see Alloy and Tabachnik 1984; Busemeyer 1991; Nisbett and Ross 1980). For example, both clinicians and laypeople overestimate the correlation between diagnostic categories (e.g., paranoia) and certain test results (e.g., unusual eye shapes in patients' drawings). This is probably because the judges' causal theories dictate a relation between the category and the result – paranoia causes patients to be especially aware of the way people look at them or of their own glances at others – since the true correlation is negligible.

Similarly, Tversky and Kahneman's (1980) experiments on causal schemas show that causal theories can dictate estimates of conditional probabilities. Participants in one experiment were asked to choose among the following options:

Which of the following events is more probable?

(a) That a girl has blue eyes if her mother has blue eyes.

(b) That a mother has blue eyes if her daughter has blue eyes.

(c) The two events are equally probable.

The converse conditional probabilities in (a) and (b) are necessarily equal, according to Bayes Theorem, provided that the (marginal or unconditional) probability of being a blue-eyed mother is the same as being a blue-eyed daughter. (A follow-up experiment verified that most participants think this equality holds.) The results showed that 45 percent of participants correctly chose option (c). The remaining participants, however, chose (a) much more often than (b): 42 percent versus 13 percent. According to Tversky and Kahneman's interpretation, these judgments are biased by the belief that it's the mother who is causally responsible for the daughter's eye color rather than the reverse. This causal asymmetry induces an incorrect impression of an asymmetry in the conditional probabilities.

Finally, Waldmann and his colleagues have shown that people's judgment about a cause can depend on causal background beliefs, even when correlational information is constant (Waldmann 1996). Consider, for example, the fictitious data in Table 2, which exhibits the relation between whether certain fruit has been irradiated and the fruit's quality in two samples, A and B. Summed over the samples, the quality of fruit is positively related to irradiation; $\Delta P$ is positive when irradiation is the factor and quality the effect. Within each sample, however, the effect reverses. Both $\Delta P$'s are negative when calculated within sample, as shown in the bottom row of the table. This situation is an example of what's known as *Simpson's paradox*: When the number of cases in the cells is unequal, the size and even the direction of contingency statistics can depend on how the population is partitioned.[9] In Table 2, people should judge irradiation to be positively related to quality if they base their decision on the entire sample, but should make the opposite judgment if they attend to samples A and B separately. Waldmann and Hagmayer (2001: Experiment 1) manipulated participants' assumptions about the causal import of the sample by informing them in one condition that sample A consisted of one type of tropical fruit and sample B consisted of a different type. In a second condition, participants learned that A and B were samples randomly assigned to two different investigators. Participants in both conditions, however, saw the same list of 80 cases (distributed as in Table 2) that identified the sample (A or B) and, for each piece of fruit, its treatment (irradiated or not) and its outcome (good or bad quality). All participants then rated how strongly irradiation affected the fruit's quality. Although correlational information was constant for the two conditions, participants rated irradiation as negatively affecting quality when the samples were causally relevant (types of fruit) but positively affecting quality when the samples were irrelevant (different investigators).

**Table 2: Contingency Information from Waldmann and Hagmayer (2001)**

| | *Sample A* | *Sample B* | *Total* |
|---|---|---|---|
| Irradiated | 16/36 | 0/4 | 16/40 |
| Not irradiated | 3/4 | 5/36 | 8/40 |
| $\Delta P$ | −.31 | −.14 | −.20 |

The first two rows indicate what fraction of a group of fruit was good as a function of whether the fruit was irradiated or not and whether it was from sample A or sample B. The top number in each fraction is the number of good fruit and the bottom number is the total number tested in that condition. Bottom row shows $\Delta P$ [i.e., Prob(Good | Irradiation) − Prob(Good-No Irradiation)] for the entire population and for each sample separately.

Given these findings, there is little chance that people construct judgments of cause from bottom up, except under the most antiseptic conditions. Naturally, this doesn't mean that contingencies, associations, and correlations are irrelevant to people's assessment of cause, but the role they play must be a piece of a much larger picture.

## POWER

As a step toward a more theory-based view of cause, we might analyze observed contingencies as due to two components: the mere

presence or absence of the cause and the tendency or power of this cause to produce the effect (Cheng 1997; Novick and Cheng 2004). The cause can't bring about the effect, of course, unless it's present. But even if it is present, the cause may be co-opted by other causes or may be too weak to produce the effect in question. Ordinarily, we can observe whether or not the cause is present, at least in the types of experiments we have been discussing, but the cause's power is unobservable. In this vein, Novick and Cheng (2004, 455) claim that "previous accounts, however, are *purely covariational* in that they do not consider the possible existence of unobservable causal structures to arrive at their output. In contrast, our theory explicitly incorporates into its inference procedure the possible existence of distal causal structures: Structures in the world that exist independently of one's observations" (emphasis in the original). On this theory, you can detect the nature of these distal structures only under special circumstances. When these special assumptions are met, the distal causal power isn't exactly an ANOVA contrast or $\Delta P$, but it looks much like a normalized $\Delta P$.

To derive the power of a cause C, suppose first that C is present in the environment. Then the effect, E, will occur in two cases: (a) C produces E (with probability $p_c$), or (b) other alternative causes, collectively designated A, occur in the same environment and produce E (with probability $\text{Prob}(A \mid C) \cdot p_a$). Thus, the probability of E when C is present is:

$$(2)\quad \text{Prob}(E \mid C) = p_c + \text{Prob}(A \mid C) \cdot p_a - p_c \cdot \text{Prob}(A \mid C) \cdot p_a.$$

The final term in (2) (after the minus sign) ensures that we count only once the case in which C and A both produce E. When C is absent, only the alternative causes A can bring about E. So the probability of E given that C is not present is:

$$(3)\quad \text{Prob}(E \mid \sim C) = \text{Prob}(A \mid \sim C) \cdot p_a.$$

Substituting these expressions in Equation (1), above, we get:

$$(4)\quad \Delta P = [p_c + \text{Prob}(A \mid C) \cdot p_a - p_c \cdot \text{Prob}(A \mid C) \cdot p_a] - [\text{Prob}(A \mid \sim C) \cdot p_a].$$

Solving (4) for $p_c$ yields the following expression for the causal power of C:

$$(5)\quad p_c = \frac{\Delta P - [\text{Prob}(A \mid C) - \text{Prob}(A \mid \sim C)]p_a}{1 - \text{Prob}(A \mid C)p_a}.$$

In the special case in which causes A and C occur independently (so that $\text{Prob}(A \mid C) = \text{Prob}(A \mid \sim C) = \text{Prob}(A)$), then Equation (5) reduces to:

$$(6)\quad p_c = \frac{\Delta P}{1 - \text{Prob}(A)p_a} = \frac{\Delta P}{1 - \text{Prob}(E \mid \sim C)}$$

The last expression follows since, by Equation (3), $\text{Prob}(E \mid \sim C)$ is equal to $\text{Prob}(A) \cdot p_a$ when A and C are independent. The interpretation of (6) may be clearer if you recall that $\Delta P$ is itself equal to $\text{Prob}(E \mid C) - \text{Prob}(E \mid \sim C)$. In other words, $p_c$ is roughly the amount that C contributes to producing E relative to the maximal amount that it could contribute. Thus, $p_c$, unlike $\Delta P$, is immune to ceiling effects – situations in which E already occurs frequently in the absence of C – except in the extreme case in which $\text{Prob}(E \mid \sim C) = 1$, where $p_c$ is undefined. (To see this, suppose $\text{Prob}(E \mid C) = .95$ and $\text{Prob}(E \mid \sim C) = .90$. Then $\Delta P = .05$, a seemingly small effect for C because both $\text{Prob}(E \mid C)$ and $\text{Prob}(E \mid \sim C)$ are high. But $p_c = .50$, a much larger effect because of the correction.) The formulas in (5) and (6) define contributory causal power, but analogous ones are available for inhibitory causal power (see Cheng 1997).

Does the power statistic, $p_c$, correspond to people's concept of a distal cause, as Novick and Cheng (2004) claim? Why shouldn't we consider it just another estimate of the likelihood that a particular cause will produce an effect – $\Delta P$ corrected for ceiling effects? The Cheng-Novick set-up portrays causation as a two-step affair. If we want to predict whether C causes E, we need to know both the likelihood that C is present and also the likelihood that C will produce E. But granting this framework, we may have some options in interpreting the latter likelihood. One issue might be whether people think that "distal power" is a probabilistic matter, as Luhmann and Ahn (2005) argue. Setting aside subatomic physics, which is outside the ken of ordinary thinking about ordinary causal interactions, people may believe that causal power is an all-or-none affair: Something either is a cause or isn't; it's not a cause with power .3 or .6. Of course, there might be reasons why a potential cause doesn't run its course, such as the failure of intermediate steps. For example, a drunk driver might have caused an accident if his car hadn't been equipped with antilock brakes. But do we want to say that the causal power of the drunk driving was some number between 0 and 1?[10] There are also cases in which a

potential cause doesn't succeed in producing its effect for reasons that we simply don't know. If we're in the dark about why a cause doesn't always produce an effect, we might want to attach a probability to it. As Cheng and Novick (2005: 703) acknowledge, "A probabilistic causal power need not indicate any violation of the power PC assumptions even for a reasoner who believes in causal determinism. . . . A probabilistic causal power might instead reflect the reasoner's imperfect representation of this cause." But this isn't consistent with Novick and Cheng's distal causal power idea. Our lack of knowledge isn't an intermediate degree of distal causal power: It's a proximal matter of our beliefs. Probabilistic beliefs about causes aren't beliefs about probabilistic causes.

Novick and Cheng are likely right that people believe that there are causes in the world and that these causes have power to produce certain effects. What's in question is whether you can model these powers as probabilities in a way that doesn't sacrifice basic intuitions about causality, which for ordinary events might be necessarily all-or-none (Luhmann and Ahn 2005) and inherently mechanistic ("intrinsically generative," in White's 2005 terms). It is possible for power proponents to retreat to the position that causal power describes an idealized, normatively correct measure that actual causal judgments merely approach. After all, distal causal powers, like distal properties and objects, are the sorts of things we infer rather than directly apprehend. However, the causal power formulas in (5) and (6), and their variants for inhibitory and interactive cases, don't necessarily yield normatively correct estimates. Like other measures of causal effectiveness – main effect contrasts, $\Delta P$, path analysis coefficients, and similar measures estimated directly from co-occurrence data – the power formulas don't always yield a normatively correct result. Glymour observes (2001: 87) that there "is an obvious reason why [the power method] will not be reliable: unobserved common causes. We have seen that the estimation methods [for generative and preventive powers] are generally insufficient when there are unobserved common causes at work, and often we have no idea before we begin inquiry whether such factors are operating." If we already know the structure of the causal environment, we can safely use power-like calculations to estimate the strength of particular pathways, and in this context, power may be a normative ideal. But this presupposes some way other than power to arrive at the correct structure.

### *Causation from Intervention*

We're finally in a position to return to the claim at the beginning of this article about "great new ideas" for representing causation. One of these ideas is the use of multiple correlations or contingencies, as in the path-analysis theories we glimpsed in the previous section. Perhaps people represent a causal system as a graph connecting causes to effects, along the lines of Figures 1 and 3. These graphs embody statistical relations – the pattern of conditional probabilities among the depicted events – that put constraints on what can be a cause of what effect. At a psychological level, we might encode this pattern of contingencies and then find the best graph – or at least a good graph – that fits them. The resulting structure is our subjective theory or causal model of the reigning causal forces. You could complain that this isn't exactly a new idea, deriving as it does from data-analytic work by Wright in the 1920s (see Wright 1960 for a recap; see also Simon 1953). But perhaps it's an innovation to take such diagrams seriously as mental representations, mental causal maps. Further elaborations may constitute genuine advances. Let's see what these could be.

We noted that graphical representations of multiple-correlation systems are open to problems of confounding and underdetermination. The very same pattern of correlations and partial correlations can be equally consistent with very different causal graphs, as the Figure 3 example illustrates. Faced with this kind of causal indeterminacy, though, scientists don't always throw up their hands. They can sometimes bring experiments to bear in selecting among the alternative causal possibilities. In the case of Figure 3, for example, imagine an experiment in which a scientist explicitly manipulates factor A to change its value. You'd expect this experiment also to change the value of B in the set up of Figure 3a but not in that of Figure 3b. Intuitively, this is because manipulating a factor can have only forward influence on its effects, not backward influence on its causes. So intervention can discriminate the two causal frameworks. Of course, we can sometimes make the same discovery without getting our hands dirty if we know the time at which the factors change their values, since causes don't follow their effects in time. In the world of Figure 3a, observing a change in A should be followed by observing a change in B, but this is not the case in Figure 3b.

Manipulating factors, however, has an advantage that goes beyond merely clarifying temporal

relationships. By changing the value of a factor, we can often remove the influence of other factors that typically covary with it, isolating the former from confoundings. If we're interested, for example, in whether listening to Mozart improves students' math scores, we could randomly assign one set of students to listen to fifteen minutes of Mozart and another to fifteen minutes of silence before a math test. In doing so we're removing the influence of intelligence, social class, and other background factors that could affect both a tendency to listen to Mozart and to do well on math tests. In the graph of Figure 3a, suppose factor A is the social class of students' families, B is intelligence, C is listening to Mozart, and D is test performance. Then the manipulation just described deletes the links from social class and intelligence to Mozart listening. In the experiment we're contemplating, students with more intelligence are no more likely to listen to Mozart than those with low intelligence. If we still find an effect of listening on test scores, this can assure us that Mozart listening affects the scores apart from the influence of the two background variables. This advantage for manipulating is due at least in part to the fact that intervention places additional constraints on the statistical relations among the variables (Pearl 2000). If we manipulate Mozart listening as just described, we're essentially creating a new graphical structure – Figure 3a minus the arrows from social class and intelligence to Mozart listening – and we're demanding that the correlations change in a way that conforms to this remodeling.

Recent evidence suggests that adults, children, and even rats are sometimes aware of the benefits of explicitly manipulating variables in learning a causal structure (Blaisdell, Sawa, Leising, and Waldmann 2006; Gopnik et al. 2004; Lagnado and Sloman 2005; Steyvers, Tenenbaum, Wagenmakers, and Blum 2003). For example, Gopnik et al. (2004) report an experiment in which four-year-olds observed a stage containing two "puppets" (simple rods with differently colored balls attached). The experimenter could move the puppets in two ways: either out of the view of the children (by reaching under the stage) or in their view (by pulling them up and down). The experimenter told the children that one of the puppets was special in that this puppet could make the other move. The children's task was to decide which was special – say, the yellow or the green one. Children first saw the yellow and green puppets moving together as the result of the experimenter's concealed action. They then observed the experimenter explicitly pulling up the yellow puppet while the green puppet remained stationary. Under these conditions, 78 percent of the children could identify the green puppet as the special one. Because a child saw the experimenter manipulate the yellow puppet without any effect on the green one, he or she could reason that the yellow puppet couldn't have been responsible for their initial joint movement and, thus, that the green puppet must be the cause. Purely association-based or correlation-based theories have trouble accounting for results like these, since such models don't distinguish between event changes that result from interventions and those that result from noninterventions.

In more complex situations (and with college-age participants), however, the advantage for interventions is not as clear-cut (Lagnado and Sloman 2005; Steyvers et al. 2003). According to Lagnado and Sloman, any benefit for intervention in their experiments seems due to the simple temporal consequences mentioned earlier (that interventions must precede their effects) rather than to the statistical independencies that interventions create. Steyvers et al. (2003: Experiment 2) presented ten observational trials about a three-variable system. They then allowed participants a single intervention, followed by an additional ten trials based on that intervention. (No explicit temporal information was available during the observation or intervention trials.) Participants' ability to identify the correct causal structure increased from 18 percent before intervention to 34 percent after (chance was 5.6 percent); however, ideal use of the intervention in this experiment should have led to 100 percent accuracy.[11] This suggests that when the environment is simple (as in Gopnik et al. 2004) and people know there are only a small number of potential causal alternatives (e.g., X causes Y vs. Y causes X), they can use facts about interventions to test which alternative is correct. When the number of alternatives is larger, hypothesis testing isn't as easy, and the participants are less able to use the difference between observations and interventions to determine the causal arrangement. Investigators have also looked at participants' ability to use a previously learned causal structure to make predictions based on observations or interventions, and we will consider the results of these experiments in the section on reasoning later in this article. The present point is that the intervention/observation difference is not

very robust when people must go from data to causal structure.

Perhaps one reason why people don't always pick up on interventions is that – as every experimentalist knows – interventions don't guarantee freedom from confounding. The literature on causal nets sometimes suggests that intervening entails only removing causal connections – links from the immediate causes of the variable that's being manipulated (i.e., the independent variable). But manipulations typically insert a new cause into the situation that substitutes for the old one in controlling the independent variable, and sometimes the new cause comes along with extraneous connections of its own. Take the example of the Mozart effect. Randomizing participants to conditions removes the influence of intelligence and other participant-centered factors. But placing participants in a control group that has to experience fifteen minutes of silence may have an aversive effect that could lower test scores to a greater extent than would merely not listening to Mozart (see Schellenberg 2005). Figuring out the right manipulation isn't always an easy matter. Ambiguity about the possible effects of an intervention may lead participants to back off from using such cues during causal learning. Of course, you can define "intervention" as a manipulation that does not affect any variable other than the one intervened on (Gopnik et al. 2004; Hausman and Woodward 1999), but this is not much help to the working scientist or layperson, who often doesn't have advance knowledge of possible side effects of the manipulation.[12]

## Reasoning from Causal Theories

We've just looked at the possibility that people discover causal relations by noticing the patterning of events in their surroundings. That method is problematic for both theoretical and empirical reasons. Theoretically, there is no limit on the number or complexity of potential causal relationships, and correlation is often unable to decide among these rival causal set ups. Empirically, there is no compelling evidence that people have hard-wired cause detectors, so people probably don't automatically derive causal facts from event perception. Moreover, our ability to infer cause from event co-occurrence seems to rely heavily on higher-level beliefs about what sorts of events can cause others, on beliefs about how events interact mechanistically, and on pragmatic pressures concerning what needs to be explained. To make matters worse, knowledge about cause sometimes colors our knowledge about co-occurrence frequency or correlation.

The classic alternative strategy for deriving causal knowledge is a form of inference to the best explanation (Harman 1965). We can start with theories about the potential causes of some phenomenon and then check to see which theory best predicts the data. The theory that provides the best fit is the one that gives the right causal picture. Of course, this form of inference doesn't give us certainty about our causal conclusions, since it depends on the range of alternatives we've considered, on the validity of the tests we've performed, and on the goodness of the data we've collected. But *no* method yields certainty about such matters. What could give us a better idea about correct causal relations than the best explanation that exploits them? This approach reserves a place for observational data, but the place is at the receiving end of a causal theory rather than at its source.

This top-down strategy, however, yields a host of further psychological problems. We still need to know the source of our theories or hypotheses if they don't arise purely from observation. We also need to consider how people use causal theories to make the sorts of predictions that hypothesis testing depends on. In this last respect, the causal schemas or Bayes nets that we looked at earlier can be helpful. We noted that people don't always accurately construct such schemes from data, even when they're allowed to manipulate relevant variables. Nevertheless, once people settle on such a representation, it may guide them to conclusions that correctly follow.

### *Representing Causal Information: Causal Principles and Causal Theories*

If we don't get causal information from innate perceptual cause detectors or from pure associative/correlational information, what's left?

#### CAUSAL PRIMITIVES

According to one top-down theory of causality, we have, perhaps innately, certain primitive causal concepts or principles that we bring to bear on the events we observe or talk about, primitives that lend the events a causal interpretation. Perhaps there is a single primitive causal relation, *cause*$(x, y)$, that we combine with other concepts to produce more complex and specific

causal descriptions (e.g., Dowty 1979; McCawley 1968; Parsons 1990). Thus, we might mentally represent the sentence in (7a) as (7b):

(7) a. John paints a picture
b. cause (John paints, become (a picture exists))

Or perhaps there are several primitive causal relations or subtypes that vary in ways that distinguish among causing, enabling, and preventing, among others (e.g., Jackendoff 1990; Schank and Riesbeck 1981; Talmy 1988; Wolff, Klettke, Ventura, and Song 2005; see also Tufte 2006 for related conclusions about causal graphs).

I suggested earlier that there was no strong evidence to support the view that people have innate cause detectors in perception, but this is consistent with the possibility of innate causal concepts. The difficulty for the perceptual view is that scenes that are supposed to trigger causal impressions automatically can usually be interpreted noncausally. But this Humean way of thinking about the perceptual demonstrations is exactly what we should expect if our interpretation of the scenes depends on how we apply our causal concepts. Having an innate concept of cause doesn't mean that external stimuli can force us to apply it. But having an innate (perceptual) cause detector – an input module in Fodor's (1983) sense – presumably does.

Of course, the existence of these concepts doesn't mean that perceptual or contingency information plays no role in our judgments about causality, and it doesn't mean that babies appear on the scene already knowing everything about causation that adults do. Percepts and contingencies can provide evidence about what we should investigate to uncover possible causal connections; however, they don't ordinarily provide a direct route to such connections. Similarly, having a causal concept may be necessary in understanding causal systems, but exactly what causes what in a particular physical setting often requires further learning. Knowing that events can be connected causally doesn't automatically tell us, for example, how chemical reactions take place or how astronomical objects interact; it simply gives us one of the ingredients or building blocks. Infants may have some domain-specific theories in areas such as psychology (Carey 1985), biology (Atran 1998), or physics (Spelke, Breinlinger, Macomber, and Jacobson 1992) that provide more specific information about causal relations in these areas, but even initial theories obviously undergo elaborations with experience and schooling, perhaps quite radical ones.

The existence of conceptually primitive causal concepts goes along with the idea that babies come equipped with the notions that events have causes, that the causes precede their effects, and that the causes bring about the effects in a mechanistic way. Bullock, Gelman, and Baillargeon (1982) propose principles along these lines – their Determinism, Priority, and Mechanism principles – and they suggest that children's and adults' later understanding of cause builds on these principles by adding information both about specific types of causal relations and about which environmental cues are most important when events interact. Preschoolers do not understand that rainbows are caused by scattering light, but they know that rainbows have some preceding mechanistic cause or other.

## CAUSAL SCHEMAS

Many cognitive theories suggest that people maintain unified representations of causal systems. If the system is the CD player in Figure 1, then memory for this information would include the individual causal relations (corresponding to the arrows in the figure) together with some larger structure that specifies how they fit together. Some theories represent the structure in terms of propositions, as in (7b), with further embedding for more complex situations (e.g., Gentner 1983); other theories employ more diagrammatic representations, similar to Figure 1 itself. The unified representations in either case may speed search for the included facts, make the included information less susceptible to interference, and highlight certain inferences. Of course, a commitment to a unified representation still leaves room for some flexibility in the representation's abstractness and completeness. It's possible that causal schemas are relatively sparse, even for familiar causal systems (Rozenblit and Keil 2002), and they may sometimes amount to little more than top-level heuristics, such as "more effort yields more results" (diSessa 2000).

As cognitive representations, causal schemas don't necessarily carry explicit information about the statistical relations among the included events. It seems possible that people could possess a schema similar to that of Figure 1 and still fail to notice the implications it has for statistical dependencies and independencies,

such as the ones we considered earlier (see the section *Causation from Correlation*). What sets Bayes nets apart from other causal schemas in psychology is their tight connection to statistical matters. Bayes nets depend essentially for their construction on a property called the (Parental) Markov condition (Pearl 2000; Spirtes, Glymour, and Scheines 2000). This is the principle that conditioning on the states of the immediate causes (the "parents") of a variable renders that variable statistically independent of all other variables in the net, except for those it causes (its "descendants"). Because the Markov principle is what determines whether a Bayes net contains or omits a link, the plausibility of Bayes nets as a psychological representation depends on the Markov condition. In the case of the CD player in Figure 1, holding constant whether the light strikes the diode will make the transmission of electrical signals independent of the rest of the variables in the figure. In the next section, we examine the empirical status of this assumption: Do people who know the causal connections in a system obey the Markov principle? In the meantime, we consider some theoretical issues that surround Bayes nets as cognitive schemas.

## CAUSAL BAYESIAN NETWORKS AND FUNCTIONAL CAUSAL MODELS AS CAUSAL SCHEMAS

Although psychologists commonly cite Pearl (2000) as a source for the theory of Bayes nets, they gloss over the fact that Pearl presents three different versions of the theory that provide successively more complex accounts of causality. These versions of Bayes nets seem to correspond to stages in the theory's evolution, with later versions placing more constraints on the representation. What Pearl refers to as "Bayesian networks" are directed graphs of variables and links that respect the Markov principle we just reviewed. What Bayesian networks depict are the pattern of statistical dependencies and independencies among a set of variables. If a set of variables X is statistically independent of another set Y given Z, then the graph displays these independencies (the graph is a *D-map* in Pearl's 1988 terminology). Conversely, if the graph displays X as independent of Y given Z, then the probability distribution contains this independency (the graph is an *I-map*). For reasons mentioned in connection with Figure 3, however, Bayesian networks do "not necessarily imply causation" (Pearl 2000: 21), since several different networks can be equally consistent with the pattern of statistical dependencies and independencies in a data set.

To overcome this indeterminacy problem, Pearl moves to a reformulated representation called "causal Bayesian networks." These networks have the same form as ordinary Bayes nets. They are still directed acyclic graphs (i.e., ones with no loops from a variable to itself), such as those in Figures 1 and 3. But causal Bayesian networks also embody constraints about interventions. These networks are answerable not just to the statistical dependencies inherent in the full graph of variables and links, but also to the statistical dependencies in the subgraphs you get when you manipulate or intervene on the variables. Within this theory, intervening on a variable means severing the connections from its parent variables and setting its value to a constant. For example, we could intervene on the "CD turns" variable in Figure 1 by disconnecting the CD holder from the motor and manually rotating it. Causal Bayes networks help eliminate the indeterminacy problem by requiring the representation to reflect all the new statistical relations that these interventions imply.

In the last part of Chapter 1 and in Chapter 7 of his book, Pearl (2000) moves to a third kind of representation: "functional causal models." At first glance, there doesn't seem to be much difference between causal Bayesian networks and functional causal models, and this might make Pearl's claims about the latter models surprising. Functional causal models are given by a set of equations of a particular type that have the form in (8):

$$(8) \quad x_i = f_i(pa_i, u_i), \quad i = 1, 2, \ldots, n.$$

Each of these equations specifies the value of one of the variables $x_i$ on the basis of the immediate (parent) causes of that variable, $pa_i$, and an additional set of variables representing other unknown factors, $u_i$, that also affect $x_i$. In the case of Figure 1, for example, we can think of the node labeled CD turns as having the value 0 if the CD is not turning and 1 if it is turning. That is, $x_{CD} = 0$ means the CD is not turning and $x_{CD} = 1$ means that it is. This value will be determined by a function like that in (8), $f_{CD}$, that will depend on the value of the parent variable (whether the motor is turning) and of a variable $u_{CD}$ (not shown in Figure 1) representing other unknown factors. Pearl considers a special case of this representation, called "Markovian causal models," in which the graph is acyclic and the $u$ terms are independent of each other, and he proves that Markovian causal models are consistent with exactly the same joint probability distributions as the corresponding causal Bayes

nets. In all probabilistic applications of Bayesian networks . . . we can use an equivalent functional model as specified in [(8)], and we can regard functional models as just another way of encoding joint distribution functions" (Pearl 2000: 31).

So what's the advantage to functional causal models that we didn't already have with causal Bayesian nets? (From now on, let's call these "causal models" and "causal nets" for short.) We noticed in discussing causal nets that the definition of these nets was given, not in terms of causal mechanisms, but in terms of probabilities. A causal net is just a Bayesian network that captures additional probability distributions, namely, the ones we get by intervening on variables. With (Markovian) causal models, we are starting in the opposite direction, beginning with functions that completely determine the states of the variables rather than beginning with probabilities. This seems consistent with the lessons of the first half of this article. As Pearl (2000: 31) puts it, ". . . agents who choose to organize their knowledge using Markovian causal models can make reliable assertions about conditional independence relations without assessing numerical probabilities – a common ability among humanoids and a useful feature for inference." Everything operates in a deterministic way in causal models, with any uncertainty confined to our lack of knowledge about the values of the $u_i$'s. Moreover, the system's equations in (8) are not just arbitrary functions that happen to give the correct $x_i$ values for cases we've observed. They reflect the actual causal determinants of the system, with $pa_i$ and $u_i$ being the true causes of $x_i$.

Pearl is explicit about the fact that an important benefit of causal models over causal networks is that the models deal correctly with counterfactual conditionals – statements of the form "If $X$ had happened, then $Y$ would have happened," like *If Fred had taken the trouble to fix his brakes, he wouldn't have had an accident.* It's been recognized at least since Goodman (1955) that there's a close connection between counterfactuals and causation. The truth of many counterfactual conditionals seems to depend on causal laws that dictate the behavior of events. These laws hold not just in our current state of affairs, but also in alternative states that differ from ours but still obey the laws in question. It's reasonable to think that the sentence about Fred is true or false because of the causal laws governing mechanical devices like brakes. If causal schemas are records of our understanding of causal laws, then they should enable us to make judgments about counterfactual conditionals. Pearl is clearly right that if causal models support counterfactuals, then this gives them a leg up on ordinary causal nets. But in order to do this, the functions in (8) have to mirror these causal laws and must be constant over all causally possible situations. Pearl outlines a specific procedure that is supposed to answer counterfactual questions ("Would $Y$ have happened if $X$ had happened?") using causal models, and we'll look at the psychological plausibility of this hypothesis in more detail in discussing causal reasoning. It's clear, though, that knowledge of causal laws (from the $f_i$'s) and knowledge of the input states of the system (from the $u_i$'s) ought to give us what we need to simulate how the system will work in all the eventualities it represents, including counterfactual ones.

The direction of explanation that Pearl's analysis takes is from causality (as given by the causal functions in (8)) to counterfactuals. At first glance, though, the opposite strategy may also seem possible. Some philosophical analyses of causation – prominently, David Lewis's (1973) – interpret causation in terms of counterfactuals. If event e would not have happened had c not happened, then e causally depends on c, according to this analysis. Psychologists have occasionally followed this lead, deciding whether one event in a story causes a second according to whether people are willing to say that the second would not have happened if the first hadn't happened (Trabasso and van den Broek 1985). Lewis's theory of counterfactuals, however, depends on similarity among possible worlds, where similarity can, in turn, depend on causal laws. The counterfactual "If c had not happened then e would not have happened" is true just in case there is a world in which neither c nor e happens that is closer to the actual world than any world where c doesn't happen but e does. And whether one world is closer to the actual world than another depends at least in part on whether the causal laws of the actual world are preserved in the alternative. Lewis didn't intend his analysis to eliminate causal laws but to provide a new way of exploiting them in dealing with relations between individual events.[13] So even if we adopt Lewis's theory, we still need the causal principles that the $f_i$'s embody (see the papers in Collins, Hall, and Paul 2004 for more recent work on the counterfactual analysis of cause).

Another possible complaint about causal models as psychological representations is that they don't come with enough structure to

explain how people are able to learn them (Tenenbaum, Griffiths, and Niyogi, in press). In figuring out how a device like a CD player works, we don't start out considering all potential networks that connect the key events or variables in the system. Instead, we take seriously only those networks that conform to our prior knowledge of what general classes of events can be causes for others. Because lasers are unlikely to turn motors, we don't waste time testing (or at least we give low weight to) causal models that incorporate such a link. According to Tenenbaum et al., people use higher-level theories to determine which network structures are possible, and this restricts the space of hypotheses they take into account. This objection seems right, since we do sometimes possess high-level knowledge (e.g., that diseases cause symptoms or that beliefs and desires cause actions) that shapes lower-level theories. Moreover, higher-level knowledge about causal laws seems necessary, given the restrictions on the $f_i$ functions that we've just discussed. But even in Tenenbaum et al.'s more elaborate hierarchy, causal models are at center-stage, mediating higher-level theory and data. This leaves us with an empirical issue: Assuming the causal models are possible psychological representations, how well do they explain people's ability to reason from their causal beliefs?

## *Causal Reasoning*

The phrase *causal reasoning* could potentially apply to nearly any type of causal thinking, including the types of causal attribution that we considered in the first part of this chapter. The issue there was how we reason *to* causal beliefs from data or other noncausal sources. Our considerations so far suggest that there may be relatively little reliable reasoning of this sort without a healthy dose of top-down causal information already in place. But how well are we able to exploit this top-down information? Once we know a batch of causal relations, how do we use them in drawing further conclusions?

### CAUSAL INTERPRETATIONS OF INDICATIVE CONDITIONALS

Cognitive psychology has tip-toed up to the issue of how people reason from causal beliefs. A number of experiments have attempted to demonstrate that inferences from conditional sentences – ones of the form *If p then q* – can depend on whether the content of the conditionals suggests a causal relation (e.g., Cummins, Lubart, Alksnis, and Rist 1991; Staudenmayer 1975; Thompson 1994). The conditionals in these experiments are indicatives, such as *If the car is out of gas, then it stalls*, rather than the counterfactual (or subjunctive) conditionals mentioned in the previous section (*If X had happened, then Y would have happened*). Because indicatives are less obviously tied to causal relationships than counterfactuals, people may reason with such conditionals in a way that does not depend on causal content.

What the results of these studies show, however, is that causal content affects people's inferences. For example, Thompson (1994) compared arguments like the ones in (9) to see how likely her participants were to say that the conclusion logically followed:

(9) a. If butter is heated, then it melts.
The butter has melted.
Was the butter heated?

b. If the car is out of gas, then it stalls.
The car has stalled.
Is the car out of gas?

Arguments (9a) and (9b) share the same form in that both have the structure: *If p then q; q; p?* So if participants attend only to this form in deciding about the arguments, they should respond in the same way to each. However, people's beliefs about cars include the fact that running out of gas is just one thing that could cause a car to stall, whereas their beliefs about butter include the fact that heating butter is virtually the only way to get it to melt. If people lean on these beliefs in determining whether the conclusions logically follow, they should be more likely to endorse the argument in (9a) than the one in (9b), and indeed they do. The difference in acceptance rates is about forty percentage points. It is possible to argue about the role played by causal information versus more abstract logical information in experiments like these, and other aspects of the data show that participants aren't simply throwing away the *if . . . then* format in favor of their causal beliefs. For our purposes, however, the question is what such experiments can tell us about the nature of those causal principles.

Thompson (1994) and others view these results as due to people's knowledge of necessary and sufficient conditions (see also Ahn and Graham 1999). Heating butter is both necessary and sufficient for its melting, whereas running out of gas is sufficient but not necessary for a car stalling. Thus, given that the butter was melted, it was probably heated; but given the car has

stalled, it may not be out of gas. The same point is sometimes made in terms of "alternative" causes or "additional" causes (e.g., Byrne 1989; Byrne, Espino, and Santamaria 1999; Cummins et al. 1991; De Neys, Schaeken, and d'Ydewalle 2003; Markovits 1984). An alternative cause is one that, independently of the stated cause (e.g., running out of gas), is able to bring about the effect, and an additional cause is one that must be conjoined with the stated cause in order for the effect to occur. The explanation of the difference between (9a) and (9b) is therefore that participants know of no alternative causes for the conditional in (9a) that would block the inference, but they do know of alternatives for the conditional in (9b) – perhaps an overheated engine or a broken fuel pump. Giving participants further premises or reminders that explicitly mention alternative or additional causes also affects the conclusions they're willing to draw (Byrne 1989; Byrne et al. 1999; De Neys et al. 2003; Hilton, Jaspars, and Clarke 1990).

The more general framing in terms of necessary and sufficient conditions, though, raises the issue of whether the experiments are tapping reasoning with specifically causal relations or with more abstract knowledge. Some of the same experiments cited earlier (Ahn and Graham 1999; Thompson 1994) demonstrate similar effects with conditionals that are about non-causal relations (e.g., conditional permissions such as *If the licensing board grants them a license, then a restaurant is allowed to sell liquor*). Likewise, you can interpret the results as due to participants' use of conditional probabilities (Evans and Over 2004; Oaksford and Chater 2003). According to Oaksford and Chater (2003), for example, people's response to the question in (9a) depends on the conditional probability that butter is heated given that it is melted, and the response to (9b) reflects the conditional probability that the car is out of gas given that it has stalled. Since the first of these is likely to be greater than the second, participants should tend to answer "yes" more often for (9a) than (9b). According to both the necessity/sufficiency and the probabilistic theories, people's beliefs about causation informs the way they represent these problems, but their reasoning is carried out over representations that don't distinguish causes from other relations.

### REASONING WITH CAUSAL VERSUS INDICATIVE CONDITIONAL STATEMENTS

We may be able to get a more direct view of how people reason about causes by looking at experiments that give participants statements containing the word *cause* or its derivatives. A number of studies have found that people make different inferences from statements of the form *p causes q* (or *q causally depends on p*) than from ones of the form *If p then q* (Rips 1983; Sloman and Lagnado 2005; Staudenmayer 1975). For example, Staudenmayer (1975) observed that participants were more likely to interpret explicit causal statements as implying a two-way, if-and-only-if, connection. For example, *Turning the switch on causes the light to go on* was more likely than *If the switch is turned on then the light goes on* to entail that the light goes on if and only if the switch is turned on. Many causal setups, however, don't lend themselves to such an interpretation. *My turning on the switch causes the light to go on* is a case in point, since the light's going on could be caused by someone else's turning. Staudenmayer included examples like these, in which the cause is not necessary for the effect. But if causal statements don't force an if-and-only-if interpretation, why the difference between causals and conditionals in the results? It seems possible that *cause* allows more freedom of interpretation than *if*. Although a two-way interpretation is possible for both *if* and *cause* in some situations (for pragmatic or other reasons), people may be more cautious about adopting it in the case of *if*.

In another respect, however, *cause* is more selective than *if*. Consider the arguments in (10):

(10) a. If the gear turns then the light flashes.
The bell rings.
Therefore, if the gear turns then both the light flashes and the bell rings.

b. The light flashing causally depends on the gear turning.
The bell rings.
Therefore, both the light flashing and the bell ringing causally depend on the gear turning.

The conclusion of (10a) seems to follow, since the conditionals are understood as statements about an existing state of affairs. The gear's turning means that the light will flash, and since the turning presumably won't affect the bell's ringing, then if the turning occurs, so will the flashing and the ringing. Argument (10a) is valid in classical propositional logic, reading *if* as the truth functional connective "$\supset$" and *and* as "&." There are many reasons to question whether natural language *if* is equivalent to $\supset$ (see Bennett 2003 for a thorough review); but even if we treat

the *if*'s in (10a) as expressing probabilistic or default relations – for example, that the conditional probability of the flashing is high given the turning, or that the turning occurs when the flashing does, all else being equal – the inference in (10a) still seems a strong one. Not so (10b). Intuitively, the conclusion asserts a causal connection between the gear's turning and the bell's ringing that goes beyond anything asserted in (10b)'s premises. In line with this impression, I found that, although 60.2 percent of participants agreed that the conclusion of arguments like (10a) had to be true whenever the premises were true, only 31.0 percent agreed to the conclusion of items like (10b) (Rips 1983). (The relatively low overall percentage of responses is probably due to the fact that the full data set included several arguments with more complex structures than that of (10).)

These differences between *cause* and *if* reflect fundamental differences in their meaning. There are disputes about the correct formal semantics for conditional sentences (see Bennett 2003). But it is plausible to think that people evaluate them by temporarily supposing that the *if*-part (antecedent) of the sentence is true and then assessing the *then*-part (consequent) in that supposed situation (Ramsey 1929/1990; Stalnaker 1968).[14] In these terms, *if* relates the current situation to a similar one (or similar ones) in which the antecedent holds. Conditionals can thus depend on circumstances that may not be a direct effect of the antecedent but simply carry over from the actual situation to the supposed one. This explains why we tend to judge that the conclusion of (10a) follows: Although the gear's turning doesn't cause the bell's ringing, nevertheless, the ringing occurs in the situation in which the gear turns. *Cause*, however, is not a sentence connective, but a predicate that connects terms for events. In order to create parallel structures between conditionals and causals in these experiments, investigators have to rephrase the antecedent and consequent as nominals (e.g., *the gear turns* in (10a) becomes *the gear turning* in (10b)), but the nominals still refer to events. Whether a causal sentence is true depends on exactly how these events are connected and not on what other circumstances may happen to hold in a situation in which the cause takes place. In this respect, causal sentences depend on the specifics of the cause-effect relation, just as ordinary predicates like *kiss* or *kick* do. Whether *John kisses Mary* is true depends on whether the appropriate relation holds between John and Mary, and whether the gear's turning causes both the light's flashing and the bell's ringing likewise depends on whether the right causal connection holds between these events. The conclusion of (10b) fails to follow from the premises, since the premises entail no such connection.

This point about the difference between conditionals and causals may be an obvious one, but analyses of *cause* can sometimes obscure it. For example, some formal treatments of action, like McCarthy and Hayes's (1969) situation calculus, represent these actions (a type of cause) as a function from a situation that obtains before the action to one that obtains after it. But although we may be able to think of both *if* and *cause* as types of functions, the truth of a causal depends more intimately on the way in which the resulting state of affairs is brought about. We judge that "if c occurs then e occurs" on the basis of whether e holds in the situations that we get by supposing c is true, but this is not enough to support the assertion that "c causes e." Similarly, there are causal modal logics (e.g., Burks 1977) that represent the causal necessity or possibility of conditionals. Such logics, for example, can symbolize sentences of the type "It is causally necessary that if c occurs then e occurs," with the interpretation that "If c occurs then e occurs" in all possible worlds that retain the actual world's causal laws. However, causally necessary conditionals aren't equivalent to causals. It is causally necessary that if $5 + 7 = 12$ then $5 + 8 = 13$, since $5 + 7 = 12$ and $5 + 8 = 13$ are true in all possible worlds, including the causally necessary ones. But $5 + 7 = 12$ doesn't cause $5 + 8 = 13$ (or anything else, for that matter), since arithmetic facts don't have causal properties.

The experiments just mentioned provide evidence that people distinguish causal sentences from indicative conditional ones, even when the conditionals have causal content. The experiments have less to say, however, about the nature of causal reasoning itself. We'd like to know in more detail how accurately people recognize inferences that follow directly from causal relations. Two possibilities present themselves, both based on our earlier discussion of causal models. First, people who know the causal facts about a system should follow the causal Markov principle in estimating probabilities of the events these models encode. Second, people's predictions about the system's behavior should respect differences between interventions and observations. We'll see that although the evidence for the first of these predictions is weak, evidence for the second is more robust.

## REASONING FROM CAUSAL MODELS: THE CAUSAL MARKOV PRINCIPLE

We've seen that Bayesian causal models (Pearl 2000) provide an explicit representation of cause-effect relations, and they include normative constraints that should govern causal reasoning. In particular, causal models obey the causal Markov principle, which provides their structural basis and mirrors statistical dependencies. We can therefore get a closer look at causal reasoning by teaching people causal connections that compose such a model and checking whether they follow the Markov principle in drawing inferences from it.

In a pioneering study of this kind, Rehder and Burnett (2005) taught participants explicit causal relations about fictional categories, such as Lake Victoria shrimp or Neptune computers. For example, participants might be told that Victoria shrimp tend to have a high quantity of ACh neurotransmitter, a long-lasting flight response, an accelerated sleep cycle, and a high body weight. The participants learned that about 75 percent of category members have each of these features. They also learned the causal relations among these features, both verbally and in an explicit diagram. For example, these participants might learn the "common cause" pattern in Figure 4a, in which high levels of ACh neurotransmitter in Lake Victoria shrimp cause a long-lasting flight response, an accelerated sleep cycle, and a high body weight. Rehder and Burnett then tested the participants by giving them descriptions of a category member with an unknown feature and asking them to rate how likely the category member was to have that feature. How likely is it, for instance, that a Victoria shrimp with high ACh, a long flight response, but no accelerated sleep cycle, also has high body weight?

The interesting predictions concern the causal Markov condition: Conditioning on the states of the parent variables renders a child variable statistically independent of all other variables, except its descendants. In the case of the figure 4a example, if we know whether a Lake Victoria shrimp has high (or low) ACh, then the values of the lower-level features – flight response and body weight, for example – will be statistically independent of each other. If we're trying to predict whether a shrimp has high body weight, it should matter a lot whether it has high or low ACh levels. But as long as we know its ACh level, we needn't worry about whether it has any of the sister features (a long flight response or an accelerated sleep cycle), since these are not descendants of body weight. It shouldn't matter how many of these sister features the shrimp has, given that it has high (low) ACh.

What Rehder and Burnett (2005) found, however, is that participants systematically violated the Markov principle. Participants' estimates of the probability that a Lake Victoria shrimp has high body weight correctly depended on whether they were told it had high levels of ACh. But these estimates also increased if the shrimp had a long flight response and an accelerated sleep cycle, even when participants knew the state of the ACh level. (See Rehder 2006; Waldmann and Hagmayer 2005: Experiment 3, for evidence of similar violations in the case of causal systems other than categories.) Rehder and Burnett's participants had learned the common-cause structure in Figure 4a, which depicts the causal model, and the Markov principle is the central ingredient in defining the model. So why do participants flagrantly disregard the principle?

Rehder and Burnett propose that participants were indeed using causal nets, but nets with a configuration that differed from the one they learned. According to this theory, the participants were assuming that there is an additional hidden node representing the category member's underlying mechanisms. The network in Figure 4b illustrates this structure, containing the new hidden mechanism node with direct connections to all the observed nodes. According to Rehder and Burnett (2005: 37), "to the extent that an exemplar has most or all of the category's characteristic features, it also will be considered a *well functioning* category member. That is, the many characteristic features are taken as a sign that the exemplar's underlying causal mechanisms functioned (and/or are continuing to function) properly or normally for members of that kind. And if the exemplar's underlying mechanisms are operating normally, then they are likely to have produced a characteristic value on the unobserved dimension." Because participants obviously aren't told the state of the hidden mechanism, the sister nodes at the bottom of the figure are no longer statistically independent. Thus, participants' tendency to rely on these sister nodes no longer violates the Markov principle. Rehder and Burnett show in further experiments that this hidden-mechanism theory also predicts the results from experiments using different network structures – for example, a net consisting of a single chain of variables and a "common effect" net with multiple causes

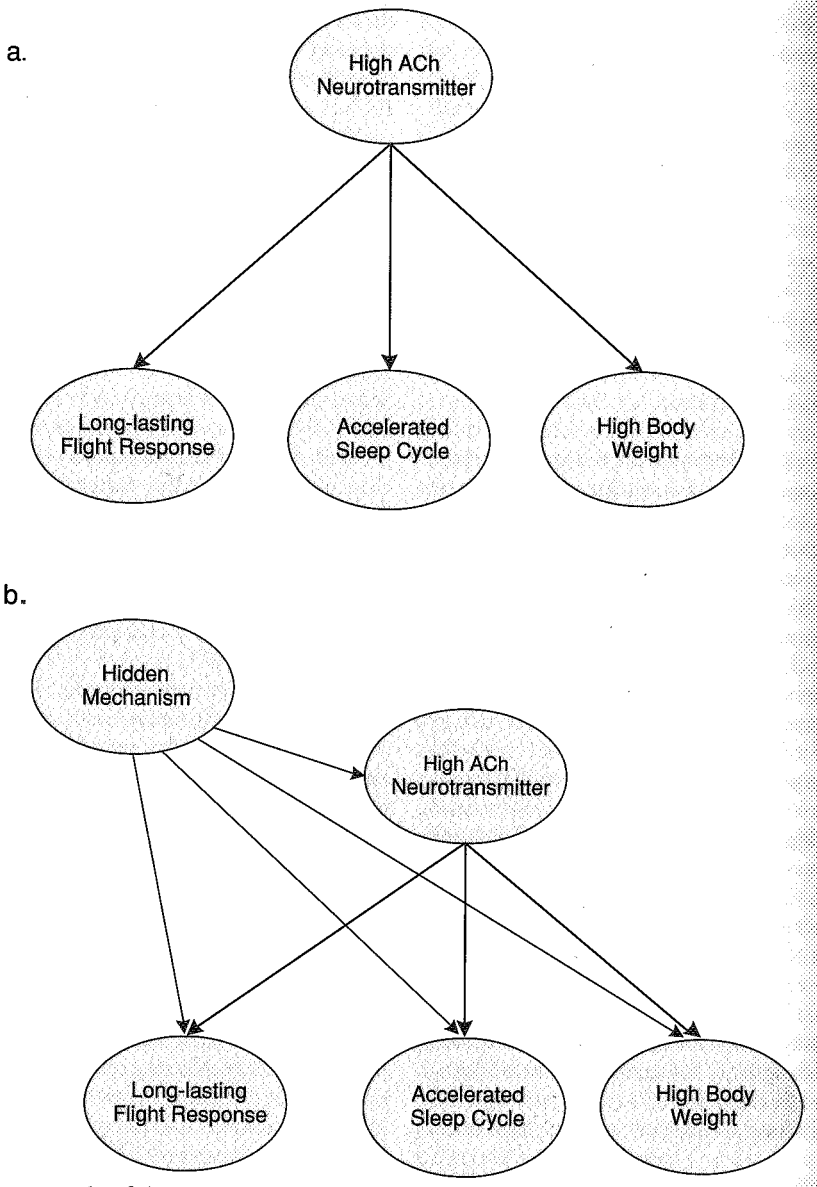


Figure 4. An example of the common cause condition from Rehder and Burnett (2005: Experiment 1). (a) The network participants learned, and (b) a possible alternative network to explain the empirical findings.

for a single effect. For the latter networks, the underlying mechanism idea seems quite plausible, and the theory is consistent with models of causal centrality and psychological essentialism (e.g., Ahn et al. 1995). Participants may suspect that a natural kind or complex artifact is likely to have some central cause or causes that hold the object together, an assumption that's in line with essentialist theories of categories (e.g., Gelman 2003; Medin and Ortony 1989). As Hausman and Woodward (1999) note, applications of the causal Markov principle have to ensure that all relevant variables are included in the model, that the causal system is analyzed at the right level, and that the included variables are not logically or definitionally related.

For common cause structures such as Figure 4a, however, why would participants go to the trouble of positing an extra hidden mechanism when they already have an explicit

common cause? Rehder and Burnett (2005: Experiment 2) also found the same pattern of results – violations of the Markov constraint – when participants were taught a common cause structure like Figure 4a for a nonsense category, daxes, whose features were arbitrarily labeled A, B, C, and D. Even if hidden mechanisms are reasonable for shrimp and computers, where you might suppose there are underlying causes in addition to those taught in the experiment, it is harder to understand why you would posit them for an obviously fictitious category. Why would participants believe there are hidden mechanisms governing well-functioning daxes? You would at least expect some decrease in the non-independence effect when the category gives participants less reason to suppose that an underlying mechanism is at work. But there doesn't seem to be much, if any, difference in the extent of the violations for daxes versus known kinds and artifacts. Although it's possible that participants were positing hidden mechanisms, a simpler alternative might be that they were reasoning in a more primitive way. Perhaps they were assuming that the dominant values of a category's features tend to cluster together, without worry too much about the exact causal set up. Participants may have been short-cutting the Bayes net circuitry, relying instead on the belief that the more typical Lake Victoria shrimp features an item has, the more likely it is to have other Lake Victoria shrimp features. Ditto for daxes. Participants weren't completely ignoring the causal structure, since they recognized the role of direct causes. But they may have given little thought to implications for the indirectly connected variables.

### REASONING FROM CAUSAL MODELS: OBSERVATION VERSUS INTERVENTION

In discussing whether people are able to infer causal nets from data (see *Causation from Intervention*), we found only limited support for the idea that people can exploit interventions in order to figure out the correct causal system. Although people use interventions within very simple systems, their ability to do so seems to fall off rapidly with even moderately complex networks. This difficulty may reflect general information-processing limits, since the number of possible causal nets (acyclic directed graphs) increases exponentially with the number of variables (see Rips and Conrad 1989). A more sensitive test of people's understanding of the intervention/observation difference is simply to give people the relevant causal relations and see whether they can predict the effects of intervening on a variable versus observing its values.

Two series of experiments provide support for sensitivity to interventions. Sloman and Lagnado (2005: Experiment 6, p. 26) gave one group of participants the problem in (11):

(11) All rocket ships have two components, A and B. Movement of Component A causes Component B to move. In other word, if A, then B. Both are moving. Suppose Component B were prevented from moving, would Component A still be moving?

A second group received the same problem, except that the final question was changed to *Suppose Component B were observed not to be moving, would Component A still be moving?* If an external process explicitly manipulates a variable – in this case, prevents Component B from moving – the internal causal connections to that variable are no longer in force, and we can't reliably use them to predict the state of the cause (Component A). By contrast, if normal internal causes are intact – if B is merely observed not to be moving – then the state of the effect provides diagnostic information about the cause. In line with this difference, 85 percent of participants responded "yes" to the intervention question, but only 22 percent did so for the observation question. A slightly more complicated problem, involving a chain of three variables instead of two, produced a similar difference between intervention questions and observation questions (Sloman and Lagnado 2005: Experiment 2). Waldmann and Hagmayer (2005: Experiment 1) also found an observation/intervention difference, using more complex five-variable systems that they presented to participants in both verbal and graphical formats.

It may seem odd, at first glance, that causal nets (or models) make correct empirical predictions in the case of the intervention/observation difference but largely incorrect predictions in the case of the causal Markov principle. This divergence might be due to differences between studies, but in fact, both results have appeared within the same experiment (Waldmann and Hagmayer 2005: Experiment 3). On second thought, though, there is no reason why these principles should necessarily hang together. We associate both the observation/intervention distinction and the Markov principle with causal nets because causal modelers have given clear formal treatments for both. And the Markov

principle, in particular, does seem tightly connected to causal nets because of the role it plays in their construction. But causal nets aren't the only way to formulate knowledge about interventions. The basic idea that you can't use the state of a manipulated variable to make inferences about its normal causes may simply be a piece of commonsense knowledge that's independent of the specific representation it gets in causal nets and models.[15] Evidence for correct understanding of interventions is support for correct causal reasoning but not necessarily support for causal nets.

## REASONING FROM CAUSAL MODELS: COUNTERFACTUALS AND CAUSE

There's one more piece of the causal net puzzle we need to consider. We've noticed substantive differences, both theoretical and empirical, between indicative conditional sentences and related causal sentences, as in (10a) and (10b). We've also noticed a much closer conceptual link between counterfactual conditionals and causals (see *Causal Bayesian Networks and Functional Causal Models as Causal Schemas*). Pearl's (2000) move from causal nets to causal models, in particular, was due to the fact that causal models give a better formulation of counterfactual questions. Causal models, but not causal nets, can tell us whether a different effect would have occurred if a cause had taken a value other than its actual one. Do causal models correctly predict people's reasoning with counterfactuals?

To handle counterfactual statements within the causal-model framework, we need a set of structural equations, like those in (8), that specify the state of each variable in terms of the state of both its parents and of uncorrelated background factors or error terms. In the simplest possible case, consider a two-variable system, such as that in (11). We can assume for the sake of this example that the all variables are dichotomous, either on or off, which we will code as 1 or 0. We can then specify the *f* functions like this:

(12) a. $f_A(u_A) = u_A$
b. $f_B(A, u_B) = A^* u_B$,

where *A* is the variable for Component A, and *B* for Component B. In other words, Component A will operate ($A = 1$) provided that the error variable, $u_A$, has the value 1, and Component B will operate ($B = 1$ ) provided both that its error variable, $u_B$, is 1 and that Component A is operating as well.

To determine the answer to a counterfactual question in this case – for example, *Suppose Component B were not operating, would Component A still operate?* – we follow a series of three steps according to Pearl (2000: Theorem 7.1.7). We first update the probability of the background *u* variables, given the current evidence about the actual state of affairs. If we assume that the two components are operating in the actual state as in (11), then $u_A = u_B = 1$. Second, we modify the causal model for an intervention on the event mentioned in the antecedent of the counterfactual. For the sample question just mentioned, we modify Component B in the usual way by orphaning B from its parent A and setting its value to a constant, while also keeping the *u* variables constant. This entails changing the equation in (12b) to:

(12) b′. $f_B(u_B) = 0$,

since the antecedent states that Component B is not operating. Finally, to determine whether Component A would still operate, we compute its probability (i.e., the value of $f_A$) in the modified model, using the updated probabilities of the background variables. Since we have $u_A = 1$, the equation in (12a) gives us a positive answer.

According to the causal model framework, the answer to our sample counterfactual question should be exactly the same as what we would get if the question had directly mentioned the manipulation of Component B. For example, we should also get a "yes" to the question: *Suppose Component B were prevented from operating, would Component A still operate?* This question is also counterfactual and differs from the first one only in making the intervention explicit. Sloman and Lagnado's (2005) Experiment 5 directly compared answers to straight counterfactuals and prevention counterfactuals, but found a reliable difference between them (68% of participants answered "yes" to the straight counterfactual and 89% "yes" to the prevention counterfactual). A similar difference appeared for scenarios describing a slightly more complicated three-variable system (Sloman and Lagnado 2005: Experiment 2). One group of participants rated the answer to a straight counterfactual (e.g., *What is the probability that A would have happened if B had not happened?*), while a second group rated an explicit prevention counterfactual (*Someone intervened directly on B, preventing it from happening. What is the probability that A would have happened?*). The average probability rating for the straight counterfactual was 3.2 on a 1–5 response scale (1 = very low, 5 = very high probability), whereas the average was 3.9 for the prevention version. Although

Sloman and Lagnado don't compare these means statistically, they do report that the first was not significantly higher than the scale midpoint (3.0), whereas the second was significantly higher than the midpoint.

Because counterfactuals were the main reason for introducing causal models (as an alternative to causal nets), it's important to see why these predictions fail. It is possible that participants are behaving in nonnormative ways in the experiments just cited, but we should also consider the possibility that the procedure itself gives an incorrect account of how counterfactuals should be understood. One thing that seems clear is that Pearl's (2000) procedure can't evaluate all reasonable counterfactuals. As he points out, the procedure is useless with "backtracking" counterfactuals that hypothesize what would have happened prior to a supposed event. For example, Sentence (13) posits an event – getting an F in a course – and gives an earlier event as a probable cause:

(13) If Fred had gotten an F in Theoretical Billiards in June, then it would have had to have been the case that he had forgotten to do his homework during the entire month of May.

Backtracking counterfactuals are sometimes clumsy to express because of tense shifts and modality, but there is no reason to think they are incoherent or uninformative. However, we would be unable to understand or evaluate backtracking counterfactuals if we had to sever the effect from its normal parent causes, since it's precisely the cause that is in question. Backtracking counterfactuals take the proposition expressed in the antecedent of the counterfactual as diagnostic of the proposition in the consequent.

Perhaps we should follow Pearl (2000) in setting aside backtracking counterfactuals and taking his procedure as a proposal about forward counterfactuals only. However, even forward counterfactuals may depend on how the hypothetical cause was brought about. Imagine that Fred's F could have been the result of two possible causes: his failure to do his homework or negligence on the part of his instructor. Then our evaluation of the truth of the forward counterfactual in (14) will depend on which of these causes we believe is the correct one:

(14) If Fred had gotten an F in the course, his instructor would have been disciplined.

As it actually happened, Fred finished his homework, his instructor was diligent, and Fred got a C. If we hold background variables constant, snip the relevant causal connections (between Fred's homework and his grade and between the instructor's behavior and his grade), and then set the grade to F, how do we determine whether the counterfactual is true or false? Intuitively, our judgment about the sentence would seem to depend on the likelihood that Fred did his homework. On one hand, if he's a marginal student, then the cause of his F is probably his own doing, and it's unlikely that the instructor will be disciplined. On the other hand, if Fred is a model student, then it may be more likely that the cause of the F was the instructor's negligence. The problem is that cutting the connection between the state of Fred's homework and his course grade renders the probability of these variables independent, and this means that the probability that his instructor will be disciplined is also independent of the homework.

The motive for cutting causal ties to the past is clear. In a deterministic system, such as those conforming to (8), no change to the actual event can occur without some alteration to its causes. To envision Fred receiving an F rather than a C, we have to envision a world in which some of the causes that produced his grade are no longer in force. We must also construct this alteration leaving as much as possible of the causal fabric of the world intact, since arbitrary changes to preceding causes give us no way to determine whether a counterfactual sentence is true or false. But although some minimal break with the past is necessary, it isn't always correct to make this break by causally isolating the event mentioned in the antecedent of the counterfactual. As the examples in (13) and (14) show, we may have to trace back to some of the causes of the antecedent event in order to see which of them is most likely to have produced the alteration. Determining which of the preceding causes must be changed may depend on which is most mutable (Kahneman and Miller 1986), as well as which is powerful enough to bring about the new effect.[16]

These reflections may help explain the differences between straight counterfactuals and prevention counterfactuals in Sloman and Lagnado's (2005) experiments. Prevention counterfactuals require explicit manipulation of the event that the antecedent of the conditionals describes. The scenario in (11) suggests that if someone had prevented Component B from operating, the intervention occurred directly at

B (perhaps by disrupting its internal mechanism). But the straight counterfactual (i.e., *Suppose Component B were not operating, would Component A still operate?*) allows more room for interpretation. We're free to imagine different ways for B to have stopped operating, some of which might plausibly involve the failure of A. Although it might seem that a world in which both A and B fail is causally more distant from the actual workaday world than one in which only B fails, this depends on details that the scenario in (11) does not supply. Stopping B by direct action on B may be more disruptive than stopping B by stopping A. There is simply no way to tell. This ambiguity is related to one we have met before in our study of causal models (in the section *Causation from Intervention*). We noted that intervening on an event means more than removing an old cause. It also entails substituting a new cause, and the way in which the intervener does this can have important consequences for what follows in the world of the intervention. The present point is that if all we know is that some event has changed from the actual situation to a counterfactual one, we have an even larger choice of mechanisms for understanding that change.

The difficulty with Pearl's (2000) account of counterfactuals doesn't mean we necessarily have to give up causal models. There may be other theories of counterfactuals based on causal schemas that provide better approaches to cases such as (13)–(14).[17] Nevertheless, people's representations of causal models are necessarily incomplete depictions of event interactions, since any event has a causal history stretching back over enormous temporal distances. We can indicate our ignorance about these prehistories by including explicit representations of uncertainty, such as Pearl's $u$ variables. But part of our causal reasoning consists in filling in some of these missing pieces, for example, in considering what sort of disturbance or manipulation could have brought about a hypothetical event. Severing preexisting connections in a model often won't be enough to explain these circumstances, since they may involve bringing in new mechanisms that we hadn't previously represented as parts of the model.

## Concluding Comments

Causal theorizing must be essential, both in everyday thinking and scientific endeavors, but it is unclear how people accomplish it. The implication of the first part of this paper is that we probably don't do such thinking by strictly bottom-up observation. We can interpret simple displays of colliding geometric shapes as instances of pushings, pullings, and other causal events. Similarly, we can interpret other swarming movements of geometrical shapes as instances of actions – for example, chasings, catchings, and fightings, as Heider and Simmel (1944) demonstrated. But we can also take a more analytical attitude to these displays, interpreting these movements as no more than approachings, touchings, and departings with no implication that one shape caused the other to move. There is no evidence to suggest that the causal interpretations are hardwired or impenetrable in the way standard perceptual illusions often are. The evidence is consistent with the idea that we see these demos as causal, but probably only in the way that we see certain visual arrays as cows or toasters. This suggestion is reinforced by the fact that, although seven-month-old infants may register some of these animations as special, others that adults report as causal are not distinctive for these infants. Of course, doubts about innate perceptual causality detectors needn't extend to doubts about innate causal concepts, but it seems likely that causal concepts, innate or learned, must have sources that aren't purely perceptual.

Are the sources of causality co-occurrence frequencies? Here there are both empirical and conceptual difficulties. On the empirical side, people are obviously limited in which potential causes they can test using frequency-based methods, and there is no theory of how they search through the space of these causes. Moreover, even when an experimenter tells participants about the relevant candidates and provides the relevant frequencies, the participants appear guided by prior hypotheses in their evaluation of the data. Theoretically, the frequency-based or correlation-based methods – main-effect contrasts, $\Delta P$, conditional $\Delta P$, Rescorla-Wagner strength, power, and path coefficients – all give incorrect answers in certain causal environments, especially when there are hidden confounding factors. Explicit manipulation or intervention can remove some of the ambiguities by eliminating the confoundings, just as in scientific experiments, but current research suggests that people are often unable to make use of such information, except in very simple settings. The empirical results are generally in line with the conclusions of Waldmann (1996) and others that

people pursue knowledge of cause in a largely top-down fashion. The theoretical results are in line with the conclusion that this might be the correct way for them to pursue it.

A top-down approach implies that people begin with hypotheses when they assess or reason about cause. But this leaves plenty of room for variation. Causal hypotheses could be anything from fragmented bits of information about a system to highly integrated and consistent theories. It's clear that people can reason with causal information and that this reasoning differs (sometimes appropriately) from what they do with similar indicative conditionals. It also seems likely that people's causal knowledge of a situation is not entirely isolated into units at the grain of individual atomic propositions (e.g., Rumelhart 1975). It is very unclear, though, what else we can say about such representations.

Bayes nets present one way of representing causal information in schematic form, and these nets provide many advantages in understanding causal situations, especially in the context of data-mining and analysis. They provide a way to factor a situation into statistically independent parts, and they therefore clarify the kinds of conclusions that we can draw from specific observations and experiments. In particular, they delimit the cases in which traditional statistical methods, such as regression or factor analysis, are likely to lead to the right results. Should we also take Bayes nets to be the mental representations that people ordinarily use to store causal facts in memory? Bayes nets go beyond a vague commitment to causal schemas in this respect, since they embody strong assumptions about the relation between the causal links in the model and statistical regularities, and they generate predictions about how people could reason about interventions and counterfactuals. They may well be consistent with the way people learn about new causal situations, though they may require additional constraints or heuristics to achieve this. In simple cases that include a small number of variables, they produce correct predictions for both children's and adults' reasoning. There seems little doubt, for example, that people observe the distinction between observation and intervention that Bayes nets embody.

On the other side of the balance, there is very little evidence that people observe the causal Markov condition, the key ingredient in Bayes net's construction. All versions of Bayes nets tie the presence and absence of causal links to the presence and absence of statistical dependencies in the data. But participants' reasoning with causal information doesn't always agree with predictions based on these dependencies. Although we can interpret the results of these experiments on the assumption that the participants are reasoning with Bayes nets that are different from the ones they are taught, there is currently little positive evidence that the Markov principle constrains people's causal reasoning. And without the Markov principle, we're back to a position not much different from ideas about cognitive schemas, models, scripts, frames, or theories that preceded Bayes nets.

Bayes nets are also oddly inarticulate as cognitive representations. Proponents of Bayes nets have generally been uninterested in the way in which people express causal regularities, presumably because people's talk about cause is filtered through pragmatic channels, obscuring their underlying beliefs. But, although this can be true, it's also the case that people's causal reasoning depends on whether a cause or set of causes is necessary or sufficient, as the literature on causal conditionals attests. Likewise, reasoning depends on the differences between independent ("alternative") and interactive ("additional") causes. While we can derive information of this sort from the underlying conditional probabilities that Bayes nets capture, we can't get them from the graphs themselves. Two arrows running into an effect could equally represent two independent, individually sufficient causes of that effect or two causes that are only jointly sufficient. The same is true for contributory versus inhibitory causes. In addition, people make a wealth of adverbial distinctions in the way that causation comes about. They distinguish, for example, between pushings, shovings, and thrustings in ways that don't seem recoverable from the bare networks or even from their underlying conditional probabilities or functional equations. These limits on expressibility may not be fundamental ones, but they do lessen the appeal of Bayes nets as cognitive maps of our causal environment.

To accord with the facts about human causal thinking, we need a representation that's less nerdy – less tied to statistical dependencies and more discursive. This doesn't mean that we should jettison Bayes nets' insights, especially insights into the differences between intervention and observation. But it does suggest that we should be looking for a representation that better highlights people's talents in describing and

reasoning about causation and downplays ties to purely quantitative phenomena.

## Notes

1 I'm grateful to Jonathan Adler, Russell Burnett, Douglas Medin, Brian Scholl, and to undergraduate and graduate students in courses on causal reasoning at Northwestern for comments on this paper.

2 Michotte (1963) is inconsistent on how to understand these reports. On the one hand, he emphasizes the phenomenal character of the observers' experiences: "Now the responses in these conditions given by the subjects always relate, of course, to the physical 'world'... But the physical 'world' in question here is no longer the world of physical *science*, as revealed by measuring instruments; it is *the world of things, as it appears to the subject* on simple inspection, his 'phenomenal world', disclosed in this case by the indications which he gives as a human 'recording instrument'. Thus, when he says that A 'pulls B' or 'pushes B', he is referring to an event occurring in a world which appears as external to him, an event of which he thinks himself simply a witness and which he is merely describing" (p. 306). But one page later, on the other hand, Michotte retreats to a position in which statements about what an observer sees are no more than abbreviations for what the observer reports: "Throughout this book there often occur expressions such as 'what the subject sees', or 'the impression received by the subject', and so on. These expressions are clearly only abbreviations, and are used to make the text less cumbersome. They in fact refer to the subjects' verbal responses and they therefore mean 'what the subject says or asserts that he sees' or 'that of which the subject says or asserts that he has an impression', and so on" (p. 307, note 5, emphasis in the original in both these passages).

3 Fodor (2003: Ch. 3) argues that even if observers directly perceive an event in the display, it's likely to be a lower-level one like square $x$ pushing another $y$ (which is indeed what observers report, according to Michotte) rather than square $x$ causing $y$ to move. There's no reason to think, according to Fodor, that perceiving an event like a pushing entails perceiving the causing. Although *x pushes y* may imply *x causes y to move*, we may get the causing from the pushing by inference rather than by direct perception. This distinction may seem unimportant to investigators, who may be satisfied that at least one type of causal interaction (pushing or launching) is directly perceived, but it is a reminder that the conclusions about direct perception have limited scope.

4 There is some debate about the exact age at which infants are first able to perceive causal interactions as such. See Cohen and Oakes (1993) for the view that infants don't fully grasp the launching interactions as causal until seven to ten months. The exact age, however, is not crucial for the issues addressed here, though the extent to which infants' recognition of causal interactions changes with experience is important. What's of interest in the present context is that infants appear to recognize oblique launching events later than linear ones.

5 For example, according to a methodology textbook by Pelham and Blanton (2003), "Most researchers who wish to understand causality rely heavily on the framework proposed by the 19th-century philosopher John Stuart Mill" (p. 63). Similarly, Cook and Campbell (1979) note, "A careful reading of chapters 3 through 7 will reveal how often a modified form of Mill's canons is used to rule out identified threats to valid inference" (p. 19). Or, in more detail, "The conditions necessary for arriving at explanations were set forth in the nineteenth century by the philosopher John Stuart Mill. . . . Mill argued that causation can be inferred if some result, $X$, follows an event, $A$, if $X$ and $A$ vary together and if it can be shown that event $A$ produces result $X$. For these conditions to be met, what Mill called the **joint method of agreement and difference** must be used. In the joint method, if $A$ occurs then so will $X$, and if $A$ does not occur, then neither will $X$" (Elmes, Kantowitz, and Roediger 1999: 103, emphasis in the original). The joint method is the third of Mill's canons, which he regarded as superior to the method of agreement but inferior to the method of difference.

6 There is also a normative problem with $\Delta P$ (as Cheng 1997 argues). Since $\Delta P$ does not take into account the presence of other causes, it can yield a misleading index of the strength of any particular cause. For example, if other causes usually bring about the effect, then $\Delta P$ for the target cause will be systematically too small. In general, measures of causal strength run into normative difficulties by ignoring the structure of the causal system (e.g., the possible presence of confounding factors). Glymour (2001) shows that this problem affects not only $\Delta P$ but also conditional $\Delta P$, Rescorla-Wagner strength, power, multiple regression coefficients, and others.

7 Chapman and Robbins (1990) and Cheng (1997) prove that under simplifying assumptions Rescorla and Wagner's (1972) theory of associative conditioning reduces to $\Delta P$. (In general, however, the equivalence does *not* hold; see Glymour 2001 citing earlier work by Danks.) A prominent member of my own faculty once declared that no graduate student from our cognitive program should get a Ph.D. without

having studied the Rescorla-Wagner model. So here's the idea: Suppose that a creature is learning a relation between a set of conditioned stimuli $C_1, C_2, \ldots, C_n$ (e.g., lights, tones, etc.) and an unconditioned stimulus $U_j$ (e.g., shock). Then the change to the associative strength, $\Delta V_i$, of a particular stimulus $C_i$ on any trial is a function of the difference between the asymptotic level of strength that's possible for the unconditioned stimulus and the sum of associative strengths for all the conditioned stimuli:

$$\Delta V_i = \alpha_i \beta_j (\lambda_j - \Sigma V_k),$$

where $\alpha_i$ is the salience of cue $C_i$, $\beta_j$ is the learning rate for $U_j$ ($0 \leq \alpha, \beta \leq 1$), $\lambda_j$ is the asymptotic level of strength possible for $U_j$, and the sum is over all cues in $C_1, C_2, \ldots, C_n$ present on the trial. The asymptote $\lambda$ will have a high value (> 0) when the unconditioned stimulus is present and a low value (perhaps 0) when it is absent on a trial. No change occurs to the strength of $C_i$ if it is not present on a trial ($\Delta V_i = 0$). The important thing to notice is that the change in strength for an individual cue depends on the strength of all others present. See Shanks and Dickinson (1987) for a discussion of the Rescorla-Wagner theory and other learning models as applied to causal judgments.

8 Psychologists tend to see ANOVA methods as superior to correlational ones in isolating the cause of some phenomenon. But as far as the statistics goes, there's no important difference between them, since ANOVA is a special case of multiple correlation/regression. The perceived difference between them is due to the fact that psychologists use ANOVA to analyze designed experiments but use correlations to analyze observational ones. Manipulation does have advantages over passive observation for reasons discussed in the following section.

9 "Simpson's paradox" is not a true paradox but an algebraic consequence of the fact that the difference between each of two proportions $a/b - c/d$ and $e/f - g/h$ can be positive (negative) while the aggregate difference $(a + e)/(b + f) - (c + g)/(d + h)$ can be negative (positive), as the numbers in Table 2 illustrate. Simpson (1951: 240) pointed out that this leaves "considerable scope for paradox and error" in how we interpret the two-way interaction between the remaining factors (i.e., the two that don't define the partition between a-d and e-h). For example, should we say that irradiation is positively or negatively related to the quality of fruit in Table 2?

10 These cases also may violate assumptions necessary in deriving $p_c$ and, if so, lie outside the domain of the power theory (see Luhmann and Ahn 2005).

11 It's also difficult to tell how much of the improvement after interventions in Stevyers et al. (2003) is due to the extra trials rather than to the interventions themselves. That is, part of participants' increased ability to identify the correct causal structure may have been the result of a larger amount of data and not to interventions per se.

12 A variation on an example of Sloman's (2005: 57–59) illustrates the same ambiguity. Suppose peptic ulcers result either from bacterial infections of a certain sort or from taking too many aspirin and similar drugs. Peptic ulcers, in turn, cause burning pains in the gut. In this situation, we may be able to intervene on someone's ulcer by administering a drug – Grandma's special formula, in Sloman's example – that cures the ulcer and thereby relieves the pain. But what should we conclude about whether the bacteria or the aspirin continue to be present after the intervention? The natural thing to say is that this depends on how Grandma's formula works. If it acts as a kind of barrier that protects the stomach lining, then perhaps the presence of the bacteria or the aspirin is unchanged. But if it works by destroying the bacteria and neutralizing the aspirin, then, of course, neither will exist after the intervention. Sloman is careful to stipulate that Grandma's special formula "goes directly to the ulcer, by-passing all normal causal pathways, and heals it every time." But how often do we know in the case of actual interventions that they route around all normal causal channels? Isn't the more usual case one where the intervention disrupts some causal paths but not others and where it may be unclear how far upstream in the causal chain the intervention takes place?

13 The old way involved deducing causal relations between individual events from general "covering" laws plus particular statements of fact (see Hempel 1965).

14 Of course, a suppositional theory needs to be worked out more carefully than can be done here. In particular, the supposition can't be such as to block all modus tollens arguments that entail the falsity of the conditional's antecedent. For a recent attempt to construct such a theory, see Evans and Over (2004).

15 This isn't to say there is no relation between the causal Markov condition and the idea of intervention. Hausman and Woodward (1999: 553) argue that "the independent disruptability of each mechanism turns out to be the flip side of the probabilistic independence of each variable conditional on its direct causes from everything other than its effects." But their argument requires a number of strong assumptions (each variable in the Bayes net must have unobserved causes and these unobserved causes can affect only one variable) that may not always be true of the representations people have of causal systems. See Cartwright (2001) for a

general critique of the causal Markov condition, and Cartwright (2002) for a specific critique of Hausman and Woodward's "flip side" claim.

16 Morteza Dehghani and Rumen Iliev have suggested factors like these in conversation.

17 In one promising account, Hiddleston (2005) proposes a causal network theory of counterfactuals that improves on Pearl (2000). Given a causal network with variables A and C, we can evaluate the truth of the counterfactual *If A = a then* C = *c* by considering all minimally different assignments of values to variables in the network such that A = a. If C = c is true in all these minimal assignments, then so is *If A = a then* C = *c*. As assignment is minimally different, roughly speaking, if (a) it has as few variables as possible whose value is different from that in the actual situation but all of whose parents have the same values, and (b) among the variables that are not effects of A, it has as many variables as possible whose values are the same as in the actual situation and all of whose parents are also the same. As Hiddleston notes, this theory allows for backtracking counterfactuals such as (13). It is unclear, however, whether this theory can capture people's intuitions about the truth of (13)–(14) and their kin. Assume a model in which turning in homework and instructor diligence are both causes of getting a grade and in which instructor diligence and the grade cause discipline of the instructor. Then there are at least two minimal models of (13)–(14) in which Fred gets an F: In one of them, Fred does his homework, the instructor is negligent, Fred gets an F, and the instructor is disciplined. In the other, Fred forgets his homework, the instructor is diligent, Fred gets an F, and the instructor is not disciplined. Since Fred does his homework in one of these models but not in the other, (13) is false, according to the theory. Similarly, for (14). As already noted, however, people's judgment of (13)–(14) may depend on how easily they can imagine the change to Fred's grade being brought about by lack of homework versus instructor negligence.

## References


Ahn, W.-K., and Graham, L. M. (1999). The impact of necessity and sufficiency in the Wason four-card selection task. *Psychological Science, 10*, 237–242.

Ahn, W.-K., Kalish, C. W., Medin, D. L., and Gelman, S. A. (1995). The role of covariation versus mechanism information in causal attribution. *Cognition, 54*, 299–352.

Ahn, W.-K., and Kim, N. (2000). The causal status effect in categorization: An overview. *Psychology of Learning and Motivation, 40*, 23–65.

Alloy, L. B., and Tabachnik, N. (1984). Assessment of covariation by humans and animals. *Psychological Review, 91*, 112–148.

Asher, H. B. (1983). *Causal modeling* (2nd ed.). Newbury Park, CA: Sage.

Atran, S. (1998). Folk biology and the anthropology of science: Cognitive universals and cultural particulars. *Behavioral and Brain Sciences, 21*, 547–609.

Barsalou, L. W. (1988). The content and organization of autobiographical memories. In U. Neisser & E. Winograd (Eds.), *Remembering reconsidered* (pp. 193–243). Cambridge: Cambridge University Press.

Barton, M. E., and Komatsu, L. K. (1989). Defining features of natural kinds and artifacts. *Journal of Psycholinguistic Research, 18*, 433–447.

Bennett, J. (2003). *A philosophical guide to conditionals*. Oxford: Oxford University Press.

Blaisdell, A. P., Sawa, K., Leising, K. J., and Waldmann, M. R. (2006). Causal reasoning in rats. *Science, 311*, 1020–1022.

Brem, S. K., and Rips, L. J. (2000). Evidence and explanation in informal argument. *Cognitive Science, 24*, 573–604.

Bullock, M., Gelman, R., and Baillargeon, R. (1982). The development of causal reasoning. In W. J. Friedman (Ed.), *The developmental psychology of time* (pp. 209–254). New York: Academic Press.

Burks, A. W. (1977). *Chance, cause, reason*. Chicago: University of Chicago Press.

Busemeyer, J. R. (1991). Intuitive statistical estimation. In N. H. Anderson (Ed.), *Contributions to information integration theory* (Vol. 1, pp. 187–215). Hillsdale, NJ: Erlbaum.

Byrne, R. M. (1989). Suppressing valid inferences with conditionals. *Cognition, 31*, 61–83.

Byrne, R. M. J., Espino, O., and Santamaria, C. (1999). Counterexamples and the suppression of inferences. *Journal of Memory and Language, 40*, 347–373.

Carey, S. (1985). *Conceptual change in childhood*. Cambridge, MA: MIT Press.

Cartwright, N. (2001). What is wrong with Bayes nets? *Monist, 84*, 242–264.

Cartwright, N. (2002). Against modularity, the causal Markov condition, and any link between the two. *British Journal for the Philosophy of Science, 53*, 411–453.

Chapman, G. B., and Robbins, S. J. (1990). Cue interaction in human contingency judgment. *Memory & Cognition, 18*, 537–545.

Chapman, L. J. (1967). Illusory correlation in observational report. *Journal of Verbal Learning and Verbal Behavior, 6*, 151–155.

Chapman, L. J., and Chapman, J. P. (1967). Genesis of popular but erroneous psychodiagnostic observations. *Journal of Abnormal Psychology, 72*, 193–204.

Cheng, P. W. (1997). From covariation to causation: A causal power theory. *Psychological Review, 104,* 367–405.

Cheng, P. W., and Novick, L. R. (1990). A probabilistic contrast model of causal induction. *Journal of Personality and Social Psychology, 58,* 545–567.

Cheng, P. W., and Novick, L. R. (1992). Covariation in natural causal induction. *Psychological Review, 99,* 365–382.

Cheng, P. W., and Novick, L. R. (2005). Constraints and nonconstraints in causal learning. *Psychological Review, 112,* 694–707.

Cliff, N. (1983). Some cautions concerning the application of causal modeling methods. *Multivariate Behavior Research, 18,* 115–128.

Cohen, L. B., and Oakes, L. M. (1993). How infants perceive a simple causal event. *Developmental Psychology, 29,* 421–433.

Collins, J., Hall, N., and Paul, L. A. (2004). *Causation and counterfactuals.* Cambridge, MA: MIT Press.

Cook, T. D., and Campbell, D. T. (1979). *Quasi-experimentation.* Chicago: Rand-McNally.

Cummins, D. D., Lubart, T., Alksnis, O., and Rist, R. (1991). Conditional reasoning and causation. *Memory & Cognition, 19,* 274–282.

De Neys, W., Schaeken, W., and d'Ydewalle, G. (2003). Inference suppression and semantic memory retrieval: Every counterexample counts. *Memory & Cognition, 31,* 581–595.

diSessa, A. A. (2000). *Changing minds: Computers, learning, and literacy.* Cambridge, MA: MIT Press.

Dowty, D. R. (1979). *Word meaning and Montague grammar.* Dordrecht, Holland: Reidel.

Einhorn, H. J., and Hogarth, R. M. (1986). Judging probable cause. *Psychological Bulletin, 99,* 3–19.

Elmes, D. G., Kantowitz, B. H., and Roediger, H. L., III. (1999). *Research methods in psychology.* Pacific Grove, CA: Brooks/Cole.

Evans, J. S. B. T., and Over, D. (2004). *If.* Oxford: Oxford University Press.

Ewert, J.-P. (1974). The neural basis of visually guided behavior. *Scientific American, 230,* 34–42.

Fodor, J. A. (1983). *Modularity of mind: An essay on faculty psychology.* Cambridge, MA: MIT Press.

Fodor, J. A. (2003). *Hume variations.* Oxford: Oxford University Press.

Gelman, S. A. (2003). *The essential child: Origins of essentialism in everyday thought.* Oxford, UK: Oxford University Press.

Gelman, S. A., and Wellman, H. M. (1991). Insides and essences: Early understanding of the non-obvious. *Cognition, 38,* 213–244.

Gentner, D. (1983). Structure mapping: A theoretical framework for analogy. *Cognitive Science, 7,* 155–170.

Gigerenzer, G. (1991). From tools to theories: A heuristic of discovery in cognitive psychology. *Psychological Review, 98,* 254–267.

Glymour, C. (2001). *The mind's arrows: Bayes nets and graphical causal models in psychology.* Cambridge, MA: MIT Press.

Goodman, N. (1955). *Fact, fiction, and forecast.* Cambridge, MA: Bobbs-Merrill.

Gopnik, A., Glymour, C., Sobel, D. M., Schulz, L., Kushnir, T., and Danks, D. (2004). A theory of causal learning in children: Causal maps and Bayes nets. *Psychological Review, 111,* 3–32.

Gregory, R. L. (1978). *Eye and brain* (3rd ed.). New York: McGraw-Hill.

Harman, G. H. (1965). The inference to the best explanation. *Philosophical Review, 74,* 88–95.

Hausman, D. M., and Woodward, J. (1999). Independence, invariance, and the causal Markov condition. *British Journal for the Philosophy of Science, 50,* 521–583.

Heider, F., and Simmel, M. (1944). An experimental study of apparent behavior. *American Journal of Psychology, 57,* 243–259.

Hempel, C. G. (1965). *Aspects of scientific explanations.* New York: Free Press.

Hiddleston, E. (2005). A causal theory of counterfactuals. *Nous, 39,* 632–657.

Hilton, D. J. (1988). Logic and causal attribution. In D. J. Hilton (Ed.), *Contemporary science and natural explanation: Commonsense conceptions of causality* (pp. 33–65). New York: New York University Press.

Hilton, D. J. (1990). Conversational processes and causal explanation. *Psychological Bulletin, 107,* 65–81.

Hilton, D. J., Jaspars, J. M. F., and Clarke, D. D. (1990). Pragmatic conditional reasoning. *Journal of Pragmatics, 14,* 791–812.

Hume, D. (1967). *A treatise of human nature* (L. A. Selby-Bigge, Ed.). Oxford: Oxford University Press. (Original work published 1739.)

Jackendoff, R. (1990). *Semantic structures.* Cambridge, MA: MIT Press.

Kahneman, D., and Miller, D. T. (1986). Norm theory: Comparing reality to its alternatives. *Psychological Review, 93,* 136–153.

Keil, F. C. (1989). *Concepts, kinds, and cognitive development.* Cambridge, MA: MIT Press.

Kelley, H. H. (1967). Attribution theory in social psychology. In D. Levine (Ed.), *Nebraska symposium on motivation* (Vol. 15, pp. 192–241). Lincoln: University of Nebraska Press.

Klem, L. (1995). Path analysis. In L. G. Grimm and P. R. Yarnold (Eds.), *Reading and understanding multivariate statistics* (pp. 65–97). Washington, DC: American Psychological Association.

Koslowski, B. (1996). *Theory and evidence: The development of scientific reasoning.* Cambridge, MA: MIT Press.

Kotovsky, L., and Baillargeon, R. (2000). Reasoning about collisions involving inert objects in 7.5-month-old infants. *Developmental Science, 3,* 344–359.

Lagnado, D. A., and Sloman, S. (2005). The advantage of timely intervention. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30*, 856–876.

Leslie, A. M. (1984). Spatiotemporal continuity and the perception of causality in infants. *Perception, 13*, 287–305.

Leslie, A. M., and Keeble, S. (1987). Do six-month-olds perceive causality? *Cognition, 25*, 265–288.

Lewis, D. (1973). Causation. *Journal of Philosophy, 70*, 556–567.

Lewis, D. (1986). Causal explanation. In *Philosophical papers* (Vol. 2, pp. 214–240). Oxford: Oxford University Press.

Loehlin, J. C. (1992). *Latent variable models: An introduction to factor, path, and structural analysis* (2nd ed.). Hillsdale, NJ: Erlbaum.

Luhmann, C. C., and Ahn, W.- K. (2005). The meaning and computation of causal power: Comment on Cheng (1997) and Novick and Cheng (2004). *Psychological Review, 112*, 685–693.

Macaulay, D. (1988). *The way things work*. Boston: Houghton Mifflin.

Markovits, H. (1984). Awareness of the "possible" as a mediator of formal thinking in conditional reasoning problems. *British Journal of Psychology, 75*, 367–376.

McCarthy, J., and Hayes, P. (1969). Some philosophical problems from the standpoint of artificial intelligence. In B. Meltzer & D. Michie (Eds.), *Machine intelligence 4* (pp. 463–502). Edinburgh: Edinburgh University Press.

McCawley, J. D. (1968). Lexical insertion in a transformational grammar without deep structure. In *Papers from the 4th regional meeting, Chicago Linguistics Society* (pp. 71–80). Chicago: Chicago Linguistics Society.

Medin, D. L., and Ortony, A. (1989). Psychological essentialism. In S. Vosniadou and A. Ortony (Eds.), *Similarity and analogical reasoning* (pp. 179–195). Cambridge: Cambridge University Press.

Michotte, A. (1963). *The perception of causality*. New York: Basic Books.

Mill, J. S. (1874). *A system of logic* (8th ed.). New York: Harper & Brothers.

Nisbett, R. L., & Ross, L. (1980). *Human inference*. Englewood Cliffs, NJ: Prentice Hall.

Novick, L. R., and Cheng, P. W. (2004). Assessing interactive causal power. *Psychological Review, 111*, 455–485.

Oakes, L. M. (1994). The development of infants' use of continuity cues in their perception of causality. *Developmental Psychology, 30*, 869–879.

Oaksford, M., & Chater, N. (2003). Conditional probability and the cognitive science of conditional reasoning. *Mind and Language, 18*, 359–379.

Parsons, T. (1990). *Events in the semantics of English: a study in subatomic semantics*. Cambridge, MA: MIT Press.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems*. San Mateo, CA: Morgan Kaufmann.

Pearl, J. (2000). *Causality*. Cambridge: Cambridge University Press.

Pelham, B. W., and Blanton, H. (2003). *Conducting research in psychology* (2nd ed). Belmont, CA: Wadsworth/Thomson.

Ramsey, F. P. (1990). General propositions and causality. In D. H. Mellor (Ed.), *Philosophical papers* (pp. 145–163). Cambridge: Cambridge University Press. (Original work published 1929.)

Rehder, B. (2006). Human deviations from normative causal reasoning. *Proceedings of the Cognitive Science Society* (p. 2596). Mahwah, NJ: Erlbaum.

Rehder, B., and Burnett, R. C. (2005). Feature inference and the causal structure of categories. *Cognitive Psychology, 50*, 264–314.

Rehder, B., and Hastie, R. (2001). Causal knowledge and categories: The effects of causal beliefs on categorization, induction, and similarity. *Journal of Experimental Psychology: General, 130*, 323–360.

Rescorla, R. A., and Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black and W. F. Prokasy (Eds.), *Classical conditioning II: Current research and theory*. New York: Appleton-Century Crofts.

Rips, L. J. (1983). Cognitive processes in propositional reasoning. *Psychological Review, 90*, 38–71.

Rips, L. J. (1989). Similarity, typicality, and categorization. In S. Vosniadou and A. Ortony (Eds.), *Similarity and analogical reasoning* (pp. 21–59). Cambridge: Cambridge University Press.

Rips, L. J. (2001). Necessity and natural categories. *Psychological Bulletin, 127*, 827–852.

Rips, L. J., and Conrad, F. G. (1989). Folk psychology of mental activities. *Psychological Review, 96*, 187–207.

Roser, M. E., Fugelsang, J. A., Dunbar, K. A., Corballis, P. M., and Gazzaniga, M. S. (2005). Dissociating processes supporting causal perception and causal inference in the brain. *Neuropsychology, 19*, 591–602.

Rozenblit, L., and Keil, F. (2002). The misunderstood limits of folk science: An illusion of explanatory depth. *Cognitive Science, 26*, 521–562.

Rumelhart, D. E. (1975). Notes on a schema for stories. In D. G. Bobrow and A. Collins (Eds.), *Representation and understanding* (pp. 211–236). New York: Academic Press.

Saxe, R., and Carey, S. (2006). The perception of causality in infancy. *Acta Psychologica, 123*, 144–165.

Schank, R. A., and Abelson, R. P. (1977). *Scripts, plans, goals, and understanding*. Hillsdale, NJ: Erlbaum.

Schank, R. A., & Riesbeck, C. K. (1981). *Inside computer understanding*. Hillsdale, NJ: Erlbaum.

Schellenberg, E. G. (2005). Music and cognitive abilities. *Current Directions in Psychological Science, 14*, 317–320.

Schlottmann, A., and Shanks, D. (1992). Evidence for a distinction between judged and perceived causality. *Quarterly Journal of Experimental Psychology, 44A*, 321–342.

Shanks, D. R., and Dickinson, A. (1987). Associative accounts of causality judgment. *Psychology of Learning and Motivation, 21*, 229–261.

Simon, H. A. (1953). Causal ordering and identifiability. In W. C. Hood and T. C. Koopmans (Eds.), *Studies in econometric method* (pp. 49–74). New York: Wiley.

Simpson, E. H. (1951). The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society, Series B, 13*, 238–241.

Sloman, S. (2005). *Causal models: How people think about the world and its alternatives*. Oxford: Oxford University Press.

Sloman, S. A., and Lagnado, D. A. (2005). Do we "do"? *Cognitive Science, 29*, 5–39.

Smith, R. H., Hilton, D. J., Kim, S. H., and Garonzik, R. (1992). Knowledge-based causal inference: Norms and the usefulness of distinctiveness. *British Journal of Social Psychology, 31*, 239–248.

Spelke, E. S., Breinlinger, K., Macomber, J., and Jacobson, K. (1992). Origins of knowledge. *Psychological Review, 99*, 605–632.

Spellman, B. A. (1996). Conditionalizing causality. *Psychology of Learning and Motivation, 34*, 167–206.

Spirtes, P., Glymour, C., and Scheines, R. (2000). *Causation, prediction, and search* (2nd ed.). Cambridge, MA: MIT Press.

Stalnaker, R. C. (1968). A theory of conditionals. In N. Rescher (Ed.), *Studies in logical theory* (pp. 98–112). Oxford: Blackwell.

Staudenmayer, H. (1975). Understanding conditional reasoning with meaningful propositions. In R. J. Falmagne (Ed.), *Reasoning: representation and process in children and adults* (pp. 55–79). Hillsdale, NJ: Erlbaum.

Steyvers, M., Tenenbaum, J. B., Wagenmakers, E., and Blum, B. (2003). Inferring causal networks from observation and interventions. *Cognitive Science, 27*, 453–489.

Talmy, L. (1988). Force dynamics in language and cognition. *Cognitive Science, 12*, 49–100.

Tenenbaum, J. B., Griffiths, T. L., and Niyogi, S. (in press). Intuitive theories as grammars for causal inference. In A. Gopnik and L. Schulz (Eds.), *Causal learning: Psychology, philosophy, and computation*. Oxford: Oxford University Press.

Thompson, V. A. (1994). Interpretational factors in conditional reasoning. *Memory & Cognition, 22*, 742–758.

Trabasso, T., & Sperry, L. (1985). Causal relatedness and importance of story events. *Journal of Memory and Language, 24*, 595–611.

Trabasso, T., and Van den Broek, P. (1985). Causal thinking and the representation of narrative events. *Journal of Memory and Language, 24*, 612–630.

Tufte, E. (2006). *Beautiful evidence*. Cheshire, CT: Graphics Press.

Tversky, A., and Kahneman, D. (1980). Causal schemas in judgment under uncertainty. In M. Fishbein (Ed.), *Progress in social psychology* (pp. 49–72). Hillsdale, NJ: Erlbaum.

Van Fraassen, B. C. (1980). *The scientific image*. Oxford: Oxford University Press.

Waldmann, M. R. (1996). Knowledge-based causal induction. *Psychology of Learning and Motivation, 34*, 47–88.

Waldmann, M. R., and Hagmayer, Y. (2001). Estimating causal strength: The role of structural knowledge and processing effort. *Cognition, 82*, 27–58.

Waldmann, M. R., and Hagmayer, Y. (2005). Seeing versus doing: two modes of accessing causal knowledge. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31*, 216–227.

Wasserman, E. A., Kao, S. F., Van Hamme, L. J., Katagiri, M., and Young, M. E. (1996). Causation and association. *Psychology of Learning and Motivation, 34*, 207–264.

Wellman, H. M. (1990). *The child's theory of mind*. Cambridge, MA: MIT Press.

White, P. A. (2005). The power PC theory and causal powers: Comments on Cheng (1997) and Novick and Cheng (2004). *Psychological Review, 112*, 675–684.

Wolff, P., Klettke, B., Ventura, T., & Song, G. (2005). Expressing causation in English and other languages. In W.-K. Ahn, R. L. Goldstone, B. C. Love, A. B. Markman, and P. Wolff (Eds.), *Categorization inside and outside the laboratory* (pp. 29–48). Washington, DC: American Psychological Association.

Wright, S. (1960). Path coefficients and path regressions: alternative or complementary concepts. *Biometrics, 16*, 189–202.