Carnegie Mellon University
Research Showcase

Department of Philosophy | Dietrich College of Humanities and Social Sciences

1-1-2004

# A Theory of Causal Learning in Children: Causal Maps and Bayes Nets.

Alison Gopnik
*University of California - Berkeley*

Clark Glymour
*Carnegie Mellon University*, cg09@andrew.cmu.edu

David M. Sobel
*Brown University*

Laura E. Schulz
*University of California - Berkeley*

Tamar Kushnir
*University of California - Berkeley*

***See next page for additional authors***

Follow this and additional works at: http://repository.cmu.edu/philosophy

Part of the Philosophy Commons

## Recommended Citation

Gopnik, Alison; Glymour, Clark; Sobel, David M.; Schulz, Laura E.; Kushnir, Tamar; and Danks, David, "A Theory of Causal Learning in Children: Causal Maps and Bayes Nets." (2004). *Department of Philosophy.* Paper 95.
http://repository.cmu.edu/philosophy/95

This Article is brought to you for free and open access by the Dietrich College of Humanities and Social Sciences at Research Showcase. It has been accepted for inclusion in Department of Philosophy by an authorized administrator of Research Showcase. For more information, please contact research-showcase@andrew.cmu.edu.

**Authors**
Alison Gopnik, Clark Glymour, David M. Sobel, Laura E. Schulz, Tamar Kushnir, and David Danks

This article is available at Research Showcase: http://repository.cmu.edu/philosophy/95

Psychological Review
2004, Vol. 111, No. 1, 3–32

Copyright 2004 by the American Psychological Association, Inc.
0033-295X/04/$12.00 DOI: 10.1037/0033-295X.111.1.3

# A Theory of Causal Learning in Children: Causal Maps and Bayes Nets

Alison Gopnik
University of California, Berkeley

Clark Glymour
Carnegie Mellon University and
Institute for Human and Machine Cognition

David M. Sobel
Brown University

Laura E. Schulz and Tamar Kushnir
University of California, Berkeley

David Danks
Carnegie Mellon University and Institute for Human and Machine Cognition

The authors outline a cognitive and computational account of causal learning in children. They propose that children use specialized cognitive systems that allow them to recover an accurate "causal map" of the world: an abstract, coherent, learned representation of the causal relations among events. This kind of knowledge can be perspicuously understood in terms of the formalism of directed graphical causal models, or Bayes nets. Children's causal learning and inference may involve computations similar to those for learning causal Bayes nets and for predicting with them. Experimental results suggest that 2- to 4-year-old children construct new causal maps and that their learning is consistent with the Bayes net formalism.

The input that reaches children from the world is concrete, particular, and limited. Yet, adults have abstract, coherent, and largely veridical representations of the world. The great epistemological question of cognitive development is how human beings get from one place to the other: How do children learn so much about the world so quickly and effortlessly? In the past 30 years, cognitive developmentalists have demonstrated that there are systematic changes in children's knowledge of the world. However, psychologists know much less about the representations that underlie that knowledge and the learning mechanisms that underlie changes in that knowledge.

In this article, we outline one type of representation and several related types of learning mechanisms that may play a particularly important role in cognitive development. The representations are of the causal structure of the world, and the learning mechanisms involve a particularly powerful type of causal inference. Causal knowledge is important for several reasons. Knowing about causal structure permits us to make wide-ranging predictions about future events. Even more important, knowing about causal structure allows us to intervene in the world to bring about new events—often events that are far removed from the interventions themselves.

Alison Gopnik, Laura E. Schulz, and Tamar Kushnir, Department of Psychology, University of California, Berkeley; Clark Glymour and David Danks, Department of Philosophy, Carnegie Mellon University, and Institute for Human and Machine Cognition, Pensacola, Florida; David M. Sobel, Department of Cognitive and Linguistic Science, Brown University.

Earlier versions and portions of this article were presented at the meetings of the International Congress on Logic, Methodology, and Philosophy of Science, Krakow, Poland, August 1999; the Rutgers Conference on the Cognitive Basis of Science, New Brunswick, NJ, September 1999; the European Society for Philosophy and Psychology meeting, Freiburg, Germany, August 2001; the Society for Philosophy and Psychology meeting, Edmonton, Alberta, Canada, June 2002; the Society for Research in Child Development meeting, Minneapolis, MN, April 2001; and the Neural Information Processing Systems meeting, Whistler, British Columbia, Canada, December 2001; and in seminars at the University of Chicago, the California Institute of Technology, Stanford University, the University of Santa Cruz, the Santa Fe Institute, and in the Cognitive Science Program and Department of Statistics at the University of California, Berkeley.

This research was supported in part by grants from the University of California, Berkeley, and the Institute of Human Development; by National Science Foundation Grant DLS0132487 to Alison Gopnik; by National Science Foundation and National Aeronautics and Space Administration Grants NCC2-1377 and NCC2-1399 to Clark Glymour; by National Institute of Health Awards F31MH12047 to David M. Sobel and F31MH66538-01A1 to Tamar Kushnir; by a National Science Foundation fellowship to Laura E. Schulz; and by McDonnell Foundation support to David Danks.

Conversations with Steve Palmer, Lucy Jacobs, Andrew Meltzoff, Josh Tenenbaum, Thierry Nazzi, John Campbell, Peter Godfrey-Smith, Henry Wellman, Peter Spirtes, John Watson, Daniel Povinelli, Stuart Russell, and the Bayes net reading group all played an important role in shaping these ideas, and we are grateful. Jitendra Malik and Marty Banks provided helpful insights and examples from vision science. Patricia Cheng made extremely helpful and thorough comments on several drafts of this article. Susan Carey also was very helpful in the revision process. We are also grateful to Christine Schaeffer and Beverly Slome for help in conducting the experiments and to the participants in our studies and their parents and teachers.

Correspondence concerning this article should be addressed to Alison Gopnik, who is at the Center for Advanced Study in the Behavioral Sciences, 75 Alta Road, Stanford, CA 94305-8090, through June 1, 2004. After June 1, 2004, correspondence should be addressed to Alison Gopnik, Department of Psychology, University of California, Berkeley, CA 94720. E-mail: gopnik@socrates.berkeley.edu

Traditionally, psychologists thought there was little causal knowledge in childhood—in particular, Piaget (1929, 1930) argued that preschoolers were precausal. In the past 2 decades, however, there has been an explosion of research on causal knowledge in young children. By the age of 5, children understand some of the basic causal principles of everyday physics (Bullock, Gelman, & Baillargeon, 1982; Leslie & Keeble, 1987; Oakes & Cohen, 1990; Spelke, Breinlinger, Macomber, & Jacobson, 1992), biology (Gelman & Wellman, 1991; Inagaki & Hatano, 1993; Kalish, 1996), and psychology (Flavell, Green, & Flavell, 1995; Gopnik & Wellman, 1994; Perner, 1991). Children as young as 2 years old can make causal predictions, provide causal explanations, and understand counterfactual causal claims (Harris, German, & Mills, 1996; Hickling & Wellman, 2001; Sobel & Gopnik, 2003; Wellman, Hickling, & Schult, 1997). Moreover, children's causal knowledge changes over time (see, e.g., Bartsch & Wellman, 1995; Gopnik & Meltzoff, 1997) and changes in light of new evidence (Slaughter & Gopnik, 1996; Slaughter, Jaakkola, & Carey, 1999). This suggests that children are actually learning about the causal structure of the world.

Much of this work has taken place in the context of the *theory theory*: the idea that children have intuitive theories of the world, analogous to scientific theories, and that these theories change in ways that are similar to scientific theory change (Carey, 1985; Gopnik, 1988; Gopnik & Meltzoff, 1997; Keil, 1989; Perner, 1991; Wellman, 1990). Causal knowledge plays a central role in theories both in science (Cartwright, 1989; Salmon, 1984) and in everyday life (Gopnik & Wellman, 1994; Gopnik & Glymour, 2002).

We argue that causal knowledge and causal learning in children involve a type of representation we call a *causal map*. Causal learning depends on learning mechanisms that allow children to recover an accurate causal map of the world. Causal maps can be inferred from observations of the patterns of correlation among events[1] or from observations of the effects of interventions—that is, actions that directly manipulate objects—or from both types of observations. We propose that young children use unconscious inductive procedures that allow them to infer causal representations of the world from patterns of events, including interventions. These procedures produce accurate representations of causal structure, at least for the most part.

We argue that these kinds of representations 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 (Pearl, 2000; Spirtes, Glymour, & Scheines, 1993, 2001). This formalism provides a natural way of representing causal structure, and it provides powerful tools for accurate prediction and effective intervention. It also provides techniques for reliably learning causal structures from patterns of evidence, including interventions. We describe the formalism, explain how it permits prediction and intervention, and describe several computational procedures by which even complicated causal maps can be learned from correlations, interventions, and prior knowledge. We suggest that children's causal learning may involve more heuristic and limited versions of similar computations. We describe experiments supporting the hypothesis that children's causal learning is in accord with Bayes net representations and learning mechanisms. These experiments also suggest that children's learning does not just involve a causal version of the Rescorla–Wagner rule.

## The Causal Inverse Problem

The study of vision has led to some of the most successful theories in cognitive science. The visual system, whether human or robotic, has to solve what has been called "the inverse problem" (Palmer, 1999). From the retinal (or foreoptic) image, the visual system has to reconstruct information about objects moving in space. Vision scientists explore how that reconstruction can be done computationally and how it is done in humans. Although accounts are very different in detail, they share some general features. (a) Visual systems have an objective problem to solve: They need to discover how three-dimensional moving objects are located in space. (b) The data available are limited in particular ways. For example, the information at the retina is two-dimensional, but the world is three-dimensional. (c) Solutions must make implicit assumptions about the spatial structure of the world and about the ways that objects in the world produce particular patterns on the retina. The system can use those assumptions to recover spatial structure from the data. In normal conditions, those assumptions lead to veridical representations of the external world. However, these assumptions are also contingent—if they are violated, then the system will generate incorrect representations of the world (as in perceptual illusions).

We propose an analogous problem about discovering the causal structure of the environment. (a) There are causal facts, as objective as facts about objects, locations, and movements, used and evidenced in accurate prediction and effective intervention. (b) The data available are limited in particular ways. Children may observe correlations between events that they cannot or do not manipulate; they may observe events they can only manipulate indirectly, through other events; the correlations they observe, with or without their own interventions, may involve an enormous number of features, only some of which are causally related. (c) Children have a causal learning system, like the visual system, that recovers causal facts by making implicit assumptions about the causal structure of the environment and the relations between the environment and evidence. Those assumptions are contingent—where they are false, causal inference may fail to get things right.

## Causal Maps

What kinds of representations might be used to solve the causal inverse problem? The visual system seems to use many very different types of representations to solve the spatial problem. But one particularly important way organisms solve the spatial inverse problem is by constructing cognitive maps of the spatial environ-

[1] We use *correlation* to signify any form of probabilistic dependence, not specifically Pearson product–moment correlation. We use ordinary capital letters (e.g., $X$, $Y$) to represent variables of any type and lowercase letters to represent their values (e.g., $x$ is a value of $X$). We use boldface letters (e.g., $\mathbf{S}$) to represent sets of variables; $X \perp Y$ denotes that for all values $x$ of $X$ and $y$ of $Y$, $X = x$ is independent in probability of $Y = y$. $W \perp \{X, Y\}|Z$ denotes that for all values $w$ of $W$, $x$ of $X$, $y$ of $Y$, and $z$ of $Z$, $W = w$ is independent in probability of both $X = x$ and $Y = y$ conditional on $Z = z$.

ment (Gallistel, 1990; O'Keefe & Nadel, 1978; Tolman, 1932). These cognitive maps provide animals with representations of the spatial relations among objects.

There are several distinctive features of cognitive maps. First, such maps provide nonegocentric representations. Animals might navigate through space, and sometimes do, egocentrically, by keeping track of the changing spatial relations between their body and objects as they move through the spatial environment. Cognitive maps are not egocentric in this way. They allow animals to represent geometric relationships among objects in space independently of their own relation to those objects. A cognitive map allows an animal that has explored a maze by one route to navigate through the maze even if it is placed in a different position initially. This aspect of cognitive maps differentiates them from the kinds of cognitive structures proposed by the behaviorists—structures that depend on associations between external stimuli and the animal's own responses.

Second, cognitive maps are coherent. Rather than just having particular representations of particular spatial relations, cognitive maps allow an animal to represent many different possible spatial relations in a generative way. An animal that knows the spatial layout of a maze can use that information to make new inferences about objects in the maze.

Third, cognitive maps are learned. Animals with the ability to construct cognitive maps can represent an extremely wide range of new spatial environments, not just one particular environment. This also means that spatial cognitive maps may be defeasible—the current representation the animal has of the environment may not be correct. As an animal explores its environment and gains more information about it, the animal alters and updates its cognitive map of that environment.

Our hypothesis is that children construct similar representations that capture the causal character of their environment. This capacity plays a crucial role in the solution to the causal inverse problem. We hypothesize that even very young children construct nonegocentric, abstract, coherent, learned representations of causal relations among events, and these representations allow them to make causal predictions and anticipate the effects of interventions.

Note that we are not proposing that children actually use spatial maps for the purpose of representing or acquiring causal knowledge or that children somehow extend spatial representations through processes of metaphor or analogy. Rather, we propose that there is a separate cognitive system with other procedures devoted to uncovering causal structure and that this system has some of the same abstract structure as the system of spatial map making.

Causal maps would be an interesting halfway point between what are traditionally thought of as domain-specific and domain-general representations. Our proposal is that these representations are specialized for causal knowledge, as spatial maps are specialized for spatial knowledge. This differentiates these representations from completely domain-general representations, such as those proposed in associationist or connectionist theories (e.g., Elman, Bates, Johnson, & Karmiloff-Smith, 1996). We predict that causal maps would not be used to represent spatial, or phonological, or musical relations, for example. However, these maps represent all kinds of causal structure. This includes the kinds of causes that are involved in everyday theories of physics, biology, and psychology as well as other kinds of causal knowledge. This differentiates these representations from the representations of more nativist modularity theories (e.g., Atran, 1990; Leslie & Roth, 1993; Spelke et al., 1992). Such theories propose that there are only a few separate domain-specific causal schemes.

Although causal maps represent causal knowledge, in particular (and in general), they are not the only devices to represent causal knowledge. Just as cognitive maps may be differentiated from other kinds of spatial cognition, causal maps may be differentiated from other kinds of causal cognition. Given the adaptive importance of causal knowledge, one might expect that a wide range of organisms would have a wide range of devices for recovering causal structure. Animals, including human beings, may have some hard-wired representations that automatically specify that particular types of events lead to other events. For example, animals may always conclude that when one object collides with another, the second object will move on a particular trajectory. Or, they may specifically avoid food that leads to nausea (Palmerino, Rusiniak, & Garcia, 1980). These sorts of specific hard-wired representations could capture particular, important parts of the causal structure of the environment. This is the proposal that Michotte (1962) and Heider (1958) made regarding the perception of physical and psychological causality.

Animals might also be hard wired to detect a wider range of causal relations that involve especially important events. Such capacities underpin classical and operant conditioning, in which animals learn associations between ecologically important events, such as food or pain, and other events. Conditioning is adaptive because it allows animals to capture particularly important causal relations in the environment.

Animals could also use a kind of egocentric causal navigation. They might calculate the immediate causal consequences of their own actions on the world and use that information to guide further action. Operant conditioning is precisely a form of such egocentric causal navigation, with special reference to ecologically important events. More generally, trial-and-error learning involves similar abilities for egocentric causal navigation.

Causal maps, however, would go beyond the devices of hard-wired representations, classical and operant conditioning, and trial-and-error learning. They would confer the same sort of advantages as spatial maps (Campbell, 1995). Most significantly, with a nonegocentric causal representation of the environment, an animal could predict the causal consequences of an action without actually having to perform it. The animal could simply observe causal interactions in the world and then produce a new action that would bring about a particular causal consequence, in the same way that an animal with a spatial map can produce a new route to reach a particular location. The capacity to produce these novel interventions would be a hallmark of causal maps.

Moreover, such an animal could combine information about the effects of its own actions, of the sort used in operant conditioning or trial-and-error learning, with purely observational information, of the sort used in classical conditioning, in a systematic way. Causal maps would also allow animals to extend their causal knowledge and learning to a wide variety of new kinds of causal relations, not just causal relations that involve rewards or punishments (as in classical or operant conditioning), not just object movements and collisions (as in the Michottean effects), and not just events that immediately result from their own actions (as in operant conditioning or trial-and-error learning). Finally, animals could combine new information and prior causal information to

create new causal maps, whether that prior information was hard wired or previously learned.

Humans, at least, do seem to have such causal representations (the case is not so clear for nonhuman animals, even including higher primates; see Povinelli, 2001; Tomasello & Call, 1997). Causal knowledge in human adults and children, particularly the sort of causal knowledge that is represented in everyday theories, seems to have much of the character of causal maps. Everyday theories represent causal relations among a wide range of objects and events in the world, independently of the relation of the observer to those objects and events (although the observer may, of course, be included in that knowledge). They postulate coherent relations among such objects and events that support a wide range of predictions and interventions, including novel interventions. These representations include a very wide range of causal facts about a very wide range of events, not just ecologically significant events. These representations go beyond the representations that would be proposed by Michottean mechanisms, classical or operant conditioning, or trial-and-error learning. Finally, the causal knowledge encoded in theories, like causal maps, appears to be learned through our experience of and interaction with the world around us.

### *Learning Causal Maps*

If the causal maps idea is correct, we can rephrase the general causal inverse problem more specifically. How is it possible to recover causal maps from the data of experience? How can we learn a new causal map? We suggest that this is one of the central cognitive problems for young children learning about the world.

The epistemological difficulties involved in recovering causal information are just as grave as those involved in recovering spatial information. Hume (1739/1978) posed the most famous of these problems, that we only directly perceive correlations between events, not their causal relationship. How can we make reliably correct inferences about whether one event caused the other? Causation is not just correlation, or contiguity in space, or priority in time, or all three, but often enough, that is our evidence.

It gets worse. Causal structures rarely just involve one event causing another. Instead, events involve many different causes interacting in complex ways. A system for recovering causal structure has to untangle the relations among those causes and discount some possible causes in favor of others.

Moreover, many causal relations may be probabilistic rather than deterministic. When a child points to a toy, this makes Mom more likely to look at the toy, but it does not mean that Mom will always look at the toy. Even if the underlying causal relationship between two kinds of events is deterministic, the occurrence of other causal factors, which may not be observed, will typically make the evidence for the relationship probabilistic. The system must be able to deal with probabilistic information.

Finally, in many cases, we make inferences about causes that are themselves unobserved or even unobservable. Something in a piece of wood makes it ignite, something in a plant makes it grow, something in a person's mind leads to action. However, we only observe the events of the wood igniting, the plant growing, or the person acting. How do we know what caused those events?

We propose that children are equipped with a causal learning system that makes certain assumptions about causal structure and about how patterns of events indicate causal structure, just as the visual system makes assumptions about spatial structure and about how retinal patterns indicate spatial structure. These assumptions help solve the causal inverse problem. Broadly speaking, there are two different kinds of assumptions that might be used to help solve the general problem of discovering causal relations.

First, we might propose what we call substantive assumptions. Children might automatically conclude that particular types of events cause other particular types of events. For example, they might assume that ingesting food causes nausea or that a ball colliding with another causes the second ball to move. The Michottean perceptual causal principles have this character, although, of course, they only allow a very limited set of causal conclusions. There might also be broader and more general assumptions of this kind, which could underpin a wider range of causal inferences: temporal sequence, for example—effects cannot precede causes. Similarly, we might propose that children automatically interpret the relation between intentional actions and the events that immediately follow those actions as causal.

Some of these substantive assumptions could be innate. But, in addition, as children learn about the world, their specific substantive knowledge about causal relations could act as a constraint on their later causal inferences. For example, if children learn that, in general, desires cause actions, they may assume that a new action was caused by a desire. Undoubtedly, substantive assumptions play an important role in solving the causal inverse problem. Innate substantive assumptions, however, would only allow children to solve a relatively limited set of causal problems with specific content. Children's capacities for causal learning appear to be much broader and more flexible than these substantive assumptions alone would allow. In the case of substantive prior causal knowledge that is not innate, there must be some other set of assumptions that allow that prior knowledge to be acquired in the first place.

We might also propose what we call formal causal assumptions. These assumptions say that certain patterns of correlation among events, including events that involve interventions, reliably indicate causal relations, regardless of the content of those events. They posit constraining relations between causal dependencies and patterns of correlations and interventions.

It is important to realize that this sort of account would not reduce causal relations between events to patterns of correlation between those events or define causal structure in terms of correlation. On our view, correlations may indicate causal structure, but they do not constitute causal structure—just as retinal patterns indicate but do not constitute spatial structure.

Our idea is that causal learning systems make certain fundamental assumptions about how patterns of correlation and intervention are related to causal relations, in much the same way that the visual system makes geometrical assumptions about how two-dimensional sensory information is related to three-dimensional space. Those assumptions may turn out to be wrong in individual cases, just as they may turn out to be wrong in the visual case. In a visual illusion, such as the illusions of depth that are produced by three-dimensional movies and Viewmaster toys, the assumptions of the visual system lead to the wrong conclusion about three-dimensional spatial structure. Similarly, on our view there might, in principle, be causal illusions, cases in which the pattern of events led to the wrong conclusion about causal structure. Overall,

and in the long run, however, these causal assumptions lead to accurate representations of the causal structure of the world. Again, as in the spatial case, this would explain why they were selected for by evolution.

Just as causal maps are an interesting halfway point between domain-specific and domain-general representations, these causal learning mechanisms are an interesting halfway point between classically nativist and empiricist approaches to learning. Traditionally, there has been a tension between restricted and domain-specific learning mechanisms such as triggering or parameter-setting and very general learning mechanisms such as association or conditioning. In the first kind of mechanism, very specific kinds of input trigger very highly structured representations. In the second kind of mechanism, any kind of input can be considered, and the representations simply match the patterns in the input. Our proposal is that causal learning mechanisms transform domain-general information about patterns of events, along with other information, into constrained and highly structured representations of causal relations.

### *Causal Maps and Causal Learning in Adults and Children*

The literature on everyday theories suggests that causal maps are in place in both adults and young children. However, there is much less evidence about the learning procedures that are used to recover those maps. For adults, there is evidence that both substantive assumptions and formal assumptions can be used to recover causal structure. Some investigators have shown that adults use substantive prior knowledge about everyday physics and psychology to make new causal judgments (e.g., Ahn, Gelman, Amsterlaw, & Kalish, 2000). Other investigators have shown that adults can also use formal assumptions to learn new causal relations—adults can use patterns of correlation among novel kinds of events to infer new causal structure. Several different specific proposals have been made to characterize such learning (e.g., Cheng, 1997; Cheng & Novick, 1992; Novick & Cheng, in press; Shanks, 1985; Shanks & Dickinson, 1987; Spellman, 1996). In particular, Cheng and her colleagues have developed the most extensive and far-reaching such account: the power PC theory (Cheng, 1997; Novick & Cheng, in press). However, adults, particularly university undergraduates, have extensive causal experience and often have explicit education in causal inference. Adults might be capable of such learning, although children are not.

There is also research showing that children and even infants use substantive assumptions and prior knowledge to make causal judgments. Most research on causal learning in children has concerned children's application of substantive principles of everyday physics. Specifically, Bullock et al. (1982) and Shultz (1982) showed that children could apply principles of everyday physics (e.g., principles involving spatial contact and temporal priority) to make new causal inferences. Work on infants also suggests that some of these principles are in place at an early age (Leslie & Keeble, 1987; Oakes & Cohen, 1990). In these experiments, children and infants seem to assume, for example, that spatial contact is required when one object causes another object to move or that causes must precede effects.

A different tradition of work suggests, at least implicitly, that children make substantive assumptions about the causal relations between their own intentional actions and the events that immediately follow those actions. For example, the literature on infant contingency learning (Rovee-Collier, 1987; Watson & Ramey, 1987) suggests that even infants can learn about the causal effects of their own actions by observing the relations between those actions and events that follow them. This learning, however, is restricted to egocentric contexts. It is analogous to trial-and-error learning in animals. The literature on imitative learning (e.g., Meltzoff, 1988a, 1988b) suggests that, at least by 9 months, infants can go beyond such purely egocentric inferences and make similar causal inferences by observing the immediate effects of the actions of others. Even in this case, however, infants seem to be restricted to considering the immediate relations between actions and the events that follow them.

However, there has been no work exploring whether young children, like adults, can use formal assumptions to recover causal maps from patterns of correlation between events. We do not even know whether children are capable of the kinds of formal causal learning that have been demonstrated in adults, let alone whether they are capable of other kinds of formal causal learning. If children can use such assumptions, that would provide them with a particularly powerful and general learning tool. Such procedures would allow children to go beyond the limited substantive knowledge that might be given innately and learn about genuinely new kinds of causal relationships and structure. Children would not only be able to infer that an object must make contact with another to cause it to move or that their own actions cause the events that follow them. Instead, they could also learn such novel causal facts as that remote controls activate television sets, that watering plants makes them grow, or that crowds make shy people nervous. Such procedures might then play a major role in the impressive changes in causal knowledge we see in the development of everyday theories. Moreover, demonstrating that this type of learning is in place in very young children would show that it does not require extended expertise or education.

### *The Role of Normative Mathematical Accounts in Psychological Research*

Do children implicitly use formal assumptions, and if so, what formal assumptions do they use? To answer this question, it would help to know which formal assumptions could, in principle, solve the causal inverse problem. Again, we may draw an analogy to vision science. Psychological solutions to the spatial inverse problem have been informed by normative mathematical and computational work. Figuring out how systems could, in principle, recover three-dimensional structure from two-dimensional information turns out to be very helpful in determining how the visual system actually does recover that information.

For example, Mayhew and Longuet-Higgins (1982) formulated a new mathematical and geometrical solution to one visual inverse problem. They showed that, in principle, depth information could be recovered from the combination of horizontal and vertical disparities between two stereo images, with no other information. (Earlier theories had argued that information about eye position was also necessary.) A three-dimensional object would generate only certain patterns of horizontal and vertical disparities and not others. As a direct result, psychophysicists tested, for the first time, whether the human visual system uses this same information in the

same way and found that, in fact, it does (Rogers & Bradshaw, 1993). In addition, computer scientists have used this mathematical account to help design computer vision systems. The mathematical theories give psychologists ways of coherently asking the question of how, and how well, humans solve these problems. Mathematical accounts of causal representations and learning could inform a psychological theory in a similar way. The Bayes net formalism provides such an account.

## Causal Bayes Nets

The causal Bayes net formalism has developed in the computer science, philosophy, and statistical literatures over the last 2 decades (Glymour & Cooper, 1999; Kiiveri & Speed, 1982; Pearl, 1988, 2000; Spirtes et al., 1993, 2001). It provides a general, unified representation of causal hypotheses that otherwise take a variety of forms as statistical models (path analysis, structural equation models, regression models, factor models, etc.). In conjunction with automated inference procedures, the formalism has been applied to design computer programs that can accurately make causal inferences in a range of scientific contexts, including epidemiology, geology, and biology (Glymour & Cooper, 1999; Ramsey, Roush, Gazis, & Glymour, 2002; Shipley, 2000). The representations of causal graphical models, commonly called Bayes nets, can model complex causal structures and generate accurate predictions and effective interventions. Moreover, Bayes net representations and associated learning algorithms can accurately infer causal structure from patterns of correlation, involving either passive observation, intervention, or both and exploiting prior knowledge. A wide range of normatively accurate causal inferences can be made, and in many circumstances, they can be made in a computationally tractable way. The Bayes net representations and algorithms make inferences about probabilistic causal relations. They also allow one to disentangle complex interactions among causes and sometimes to uncover hidden unobserved causes (see Glymour, 2001; Glymour & Cooper, 1999; Jordan, 1998; Pearl, 1988, 2000; Spirtes et al., 1993, 2001).

### *Inferring Causal Structure From Conditional Dependence: An Informal Example*

Bayes nets are actually a formalization, elaboration, and generalization of a much simpler and more familiar kind of causal inference. Causal relations in the world lead to certain characteristic patterns of events. If *X* causes *Y*, the occurrence of *X* makes it more likely that *Y* will occur. We might think that this could provide us with a way of solving the causal inverse problem. When we see that *X* is usually followed by *Y*, we can conclude that *X* caused *Y*.

But there is a problem. The problem is that other events might also be causally related to *Y*. For example, some other event, *Z*, might be a common cause of both *X* and *Y*. *X* does not cause *Y*, but whenever Z occurs, it is more likely that both *X* and *Y* will occur together. Suppose you notice that when you drink wine in the evenings, you are likely to have trouble sleeping. It could be that the wine is causing your insomnia. However, it could also be that you usually drink wine in the evenings when you go to a party. The excitement of the party might be keeping you awake, independently of the wine. The party might both cause you to drink wine and independently cause you to be insomniac, and this might be responsible for the association between the two kinds of events. *X* (wine) would be associated with *Y* (insomnia), yet it would be wrong to conclude that there was a causal relation between them. You could represent these two possibilities with two simple graphs:

1. *Z* (Parties) $\rightarrow$ *X* (Wine) $\rightarrow$ *Y* (Insomnia). (Parties cause wine drinking, which causes insomnia.)

2. *X* (Wine) $\leftarrow$ *Z* (Parties) $\rightarrow$ *Y* (Insomnia). (Parties cause wine drinking and also cause insomnia.)

Intuitively, these two causal structures should lead to different patterns of correlation among the three types of events. If Graph 1 is right, you can predict that you will be more likely to have insomnia when you drink wine, regardless of how much you party. If Graph 2 is right, you can predict that you will be more likely to have insomnia when you party, regardless of how much you drink. Similarly, different interventions will be adaptive in these two cases. If Graph 1 is right, then you should avoid drinking to help cure your insomnia (even solitary drinking), and if Graph 2 is right, you should avoid parties (even sober parties).

According to either of these graphs, drinking wine is correlated with insomnia, and the fact that you are drinking wine increases the probability that you will have insomnia, but for two different reasons. In Graph 1 the two events are related because wine drinking causes insomnia, but in Graph 2 they are related because the fact that you are drinking wine increases the probability that you are at a party, and parties cause insomnia. If Graph 1 is true the correlation between wine drinking and insomnia tracks the probability that an intervention to stop drinking will reduce insomnia. But if Graph 2 is true, the correlation between wine drinking and insomnia does not track this probability—not drinking will not help. In much the same way, smoking is correlated both with having yellow fingers and with getting cancer, so having yellow fingers is correlated with cancer, but cleaning your hands will not keep you from getting cancer, and quitting smoking will. Knowing the right causal structure may not be essential for predicting one thing from another, but it is essential for predicting the effects of interventions that deliberately manipulate events.

If you knew which of these graphs was right, you could manage to get some sleep without unnecessarily sacrificing alcohol or society. How could you decide? Intuitively, it seems that you could work backward from knowing how the graphs lead to patterns of events to inferring the graphs from those patterns of events. One thing you could do is to perform a series of experimental interventions, holding wine or parties constant, and varying the other variable. Because you already know that social drinking is associated with insomnia, you could systematically try solitary drinking or sober partying and observe the effects of each of these interventions on your insomnia.

You could also, however, simply collect observations of the relative frequencies of *X*, *Y*, and *Z*. If you observe that you are more likely to have insomnia when you drink wine, whether or not you are at a party, you could conclude that the wine is the problem. If you observe that you are more likely to have insomnia only when you go to a party, regardless of how much or how little wine you drink, you could conclude that the parties are the problem. In

both cases, wine, insomnia, and partying will all be correlated with one another. But if Graph 1 is correct, then insomnia will continue to be dependent on wine even if you take partying into account—insomnia is still dependent on wine conditional on partying. However, insomnia will no longer be dependent on partying if you take wine into account—insomnia and partying are independent conditional on wine. In contrast, if Graph 2 is correct, then insomnia and partying will still be dependent if you take wine into account—insomnia and partying are dependent conditional on wine. However, insomnia will no longer be dependent on wine when you take partying into account—insomnia and wine are independent conditional on partying. In this simple case, then, you have figured out which of two causal structures is correct by observing the patterns of conditional dependence and independence among events.

This sort of reasoning is ubiquitous in science. In experimental design, scientists control for events that might be confounding causes. In observational studies, they use techniques such as partial correlation to control for confounding causes. In effect, what you did in your reasoning about your insomnia was to design an experiment controlling for partying or to partial out the effects of partying from the wine–insomnia correlation.

We can translate these informal intuitions about conditional dependence into the more precise language of probability theory (see Reichenbach, 1956). More formally, we could say that if Graph 1 is right and there is a causal chain that goes from parties to wine to insomnia, then $Y \perp Z|X$—the probability of insomnia occurring is independent (in probability) of the probability of party going occurring conditional on the occurrence of wine drinking. If Graph 2 is right and parties are a common cause of wine and insomnia, then $X \perp Y|Z$—the probability of wine drinking occurring is independent (in probability) of the probability of insomnia occurring conditional on the occurrence of party going.

There is also a third basic type of graph; insomnia might be a common effect of both wine drinking and parties ($X \rightarrow Y \leftarrow Z$). In this case, $X$ is not independent of $Z$ conditional on $Y$. The intuitions here are less obvious, but they reflect the fact that, in this case, knowing about the effect and about one possible cause gives us information about the other possible cause. We can illustrate this best with a different example. Suppose $X$ is a burglar, $Y$ is the burglar alarm sounding, and $Z$ is the neighboring cat tripping the alarm wire, so that $Y$ is a common effect of X and $Z$. If we hear the alarm sound and see the cat tripping the wire, we are less likely to conclude that there was a burglar than if we simply hear the alarm sound by itself (see Pearl, 2000; Spirtes et al., 1993, for discussion).

The notions of dependence and independence involved in these formal statements, unlike the intuitive notions, are precisely defined in terms of the probabilities of various values of the variables. Suppose each of the variables in our example has only two values: People drink, party, or have insomnia, or they do not. If drinking and insomnia are independent in probability, then the probability of drinking and insomnia occurring together equals the probability of drinking occurring multiplied by the probability of insomnia occurring. Similarly, the probability that drinking does not occur and that insomnia also does not occur equal the probability of not drinking multiplied by the probability of not having insomnia. More generally, the same is true for any combination of values of $X$ and $Y$.

However, if $X$ and $Y$ are dependent in probability, then there will be some set of values of $X$ and $Y$ such the probability of their occurring together will not equal the probability of $X$ occurring multiplied by the probability of $Y$ occurring. For example, if drinking wine causes insomnia, then the probability of drinking and insomnia occurring together will be greater than the probability of drinking occurring multiplied by the probability of insomnia occurring.

Formally, $X$ and $Y$ are independent in probability if and only if for every value of $X$ and $Y$

$$Pr(X, Y) = Pr(X) \times Pr(Y).$$

Similarly, we can define the conditional independence of two variables given another value in these probabilistic terms. $X$ (drinking) is independent of $Y$ (insomnia) conditional on $Z$ (parties) if and only if for every value of $X$, $Y$, and $Z$[2]

$$Pr(X, Y|Z) = Pr(X|Z) \times Pr(Y|Z).$$

The informal reasoning we described above is limited to rather simple cases. But, of course, events may involve causal interactions among dozens of variables rather than just three. The relations among variables may also be much more complicated. $X$ might be linearly related to $Y$, or there might be other, more complicated functions relating $X$ and $Y$; $X$ might inhibit $Y$ rather than facilitating it, or $X$ and $Z$ together might cause $Y$ though neither event would have that effect by itself. And, finally, there might be other, unobserved, hidden variables that are responsible for patterns of correlation. Is there a way to take the probability theoretic statement of the reasoning we use intuitively and generalize it to these more complicated cases? The causal Bayes net formalism provides such a method.

### *Bayes Nets*

The causal Bayes net formalism has three aspects: Directed acyclic graphs represent causal relations, the graphs are associated with probability distributions, and the Markov assumption constrains those probability distributions. We state these three aspects more formally first and then give a more informal and discursive explanation.

1. Causal hypotheses are represented by directed acyclic graphs, in which the causal variables or features are the nodes or vertices. A directed edge between two variables, $X \rightarrow Y$, stands for the proposition that there is some intervention that fixes an assignment of values to all other variables represented in the graph (resulting in $Y$ having a particular probability distribution $pr[Y]$) such that an intervention that (a) changes the value of $X$ from $x$ to some $x'$ for distinct values $x$, $x'$ of $X$ but (b) does not influence $Y$ other than

[2] Independence and conditional independence can also be defined in other ways. For variables $X$, $Y$, $Z$ taking only two possible values (denoted, e.g., by $Y$ and $\sim Y$ and bearing in mind that $Y$ is ambiguously a variable and a value of that variable), that $X$ is independent of $Y$ conditional on $Z$ can also be defined by a "difference" formula, $Pr(X|Y, Z) = Pr(X|\sim Y, Z)$, for all values of $X$, $Y$, and $Z$. The two definitions are equivalent when all represented conditional probabilities exist, but not necessarily otherwise. For technical reasons, the definition in the body of the text is customary in the Bayes net formalism.

through $X$ and (c) does not change the fixed values of other variables will result in a probability distribution $pr'(Y) \neq pr(Y)$ (see Spirtes et al., 1993, for a full mathematical justification of this characterization).

2. There is a joint probability distribution on all assignments of values to all variables in the graph. Typically, a particular probability distribution can be specified by values of parameters that may be linked to the structure of the graph, with sets of different parameter values specifying different probability distributions. For example, in a linear system with a normal (Gaussian) distribution, in which each variable is a linear function of its direct causes and of unobserved factors (often described as noise or error), the parameters can be given by the linear coefficients and the covariance matrix of the unobserved factors. Often (in regression models, e.g.) the covariance matrix is assumed to be diagonal—that is, the unobserved factors are assumed to be independent in probability. The directed acyclic graph, the numerical values for the linear coefficients, the variance of the unobserved factors, and the specification that the unobserved factors are jointly independent in probability determine a unique joint probability distribution on all values of the variables. If the variables are discrete—for example, if they take only two possible values, for instance, *present* or *absent*—the parameters are simply the probability distributions for each variable conditional on each possible assignment of values to its direct causes in the graph.

3. The joint probability distribution on all assignments of values to all variables in the graph is constrained in the following way: For any variable, $R$, in the directed graph, the graph represents the proposition that for any set, **S**, of variables in the graph (not containing any descendants of $R$), $R$ is jointly independent of the variables in **S** conditional on any set of values of the variables that are parents of $R$ (the direct causes of $R$, those variables that have edges directed into $R$). In the Bayes net literature, this condition is called the Markov assumption. If the marginal probability distributions of each variable in a directed acyclic graph conditional on each vector of values of its parent variables are a function of the values of its parents alone, the Markov assumption necessarily follows.

Causal Bayes nets, then, represent causal structure in terms of directed graphs, like the simple graphs we used in the wine–insomnia example or the more complex graph shown below in Figure 1. The nodes of the graph represent variables whose values are features of the system to which the net applies. Color, for example, might be a variable with many different possible values; weight might be a variable with a continuum of values; having eyes might be a variable with just two discrete values, *absent* or *present*.

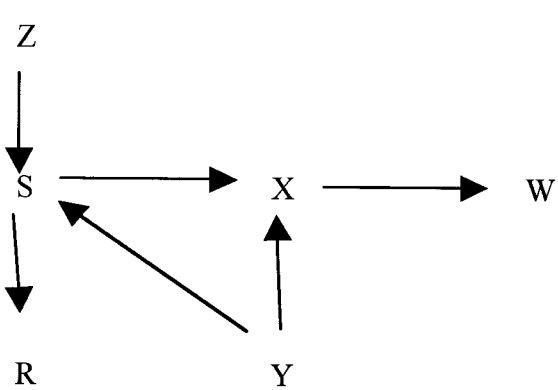


*Figure 1.* A causal graph.

In a causal Bayes net, the arrows represent the direct causal relations between two variables. These causal relations are objective relations among types of objects and events in the world, the sorts of relations scientists discover. There are, of course, knotty philosophical questions about the metaphysics of these causal relations. But, at the least, we can assume that these causal facts lead to facts about the effects of interventions on the world—indeed, this is why science is possible.

From this perspective, the precise definition of an arrow between $X$ and $Y$ given in Point 3 of the formal characterization above can be roughly translated as follows: If we did the right experiment, controlling all the other variables in the graph, changing the value of $X$ would directly cause a change in the value of $Y$. Similarly, in Figure 1, for example, if we fixed the value of $X$, $Y$, $W$, and $Z$ and then changed the value of $S$, the value of $R$ would change.

The complex definition in the three parts of Clause 1 above is essential for generality, exactly because the correlations among a set of variables do not uniquely determine their causal relations. For example, $X \rightarrow Y \rightarrow Z$ and $X \leftarrow Y \leftarrow Z$ are distinct causal hypotheses, but they imply the same constraint on probabilities: $X$, $Y$, and $Z$ are correlated, and $X$ is independent of $Z$ conditional on $Y$. However, these hypotheses imply different predictions about the result of (ideal) interventions. For example, only the first hypothesis implies that interventions that alter $X$ while fixing $Z$ at a constant value throughout will alter the probability of $Y$. Details are given in Chapters 3 and 4 of Spirtes et al. (1993, 2001).

These directed graphs must also be acyclic. An arrow going out from one node cannot also lead back into that node (e.g., $X \rightarrow Y \rightarrow Z \rightarrow X$). No feedback loops are allowed in the graphs (although there are generalizations to cyclic graphs; see Richardson, 1996, and Spirtes et al., 2001).

These graphs may encode deterministic relations among the variables (so, e.g., $S$ might always lead to $X$). More often, the causal relations among the variables are conceived of as probabilistic (either because they are intrinsically probabilistic or because there is assumed to be unmeasured noise due to variations in other unrepresented causes). So, for example, $S$ might have a .78 probability of leading to $X$. The relations may also vary in other ways—for example, they might be inhibitory, or additive, or linear, or nonlinear. The parameterization of a graph provides additional information about the statistics of the causal relations (e.g., whether they are deterministic, or probabilistic, or linear or nonlinear). This information goes beyond the information that $S$ directly causes $X$, which is encoded by the arrow itself.

The Markov assumption says that, if the graph is causal, there are certain conditional independence relations among the variables, no matter how the graph is parameterized, and it defines those relations. (The Markov assumption does not characterize conditional independence relations that hold only for particular parameter values.) We can use kinship terms to characterize various relations among the arrows and help explain the Markov assumption more intuitively. Thus, if, as in Figure 1, $S$ and $X$ are directly connected by an arrow that is directed into $X$, $S$ is a parent

of $X$ and $X$ is a child of $S$. Similarly, we can talk about ancestors and descendants to characterize indirect relations among the variables. In Figure 1, $Z$ is an ancestor of $X$ and $X$ is a descendant of $Z$. The Markov assumption says that the variables in a causal network are independent of all other variables in the network, except their descendants, conditional on their parents. For example, in Figure 1, the Markov assumption says that $X$ is independent of $\{R, Z\}$ conditional on any values of variables in the set $\{S, Y\}$.

In general, the Bayes net formalism allows us to take information about the correlations of some variables in a causal network and/or about the results of experimental interventions on some variables and then correctly infer the correlations among other variables and/or the results of experimental intervention on those variables. The arrows encode propositions about the effects of interventions on a variable, and from those arrows we can make new inferences about other correlations and interventions, as we show below in the section on prediction and planning. Conversely, we can infer the arrows—that is, infer propositions about interventions on some variables—from information about the correlations among other variables and about the effects of interventions on other variables, as we show below in the section on learning. And, as we also show, in some cases we can do this even when the variables are not observed. No matter what the metaphysics of causation may be, these sorts of inferences are central to causal learning.

### *Using Causal Bayes Nets for Prediction and Planning*

A directed acyclic graph, with a probability distribution constrained by the Markov assumption, represents a set of causal hypotheses about a system. Given such a graph, we can make two kinds of normatively accurate inferences. First, we can use the graph to predict the value of a variable in the system from observed values of other variables in the system. However, we can also make another type of inference, often quite different from the first. We can predict the value of a variable when actions intervene from outside the system to directly alter the values of other variables. The causal Bayes net formalism provides algorithms for both kinds of inference.

*Prediction.* The first prediction problem is this: Given a Bayes net (i.e., a directed acyclic graph and associated probability distribution that obeys the Markov assumption), given any variable $X$ in the graph, and given any vector $\mathbf{V}$ of values for any set $\mathbf{S}$ of other variables in the graph, compute the probability of $X$ conditional on the values $\mathbf{V}$ for the variables in $\mathbf{S}$. Bayes nets were originally applied to solve this problem in expert systems in artificial intelligence. They were first used to help calculate the conditional probabilities among sets of variables. A variety of efficient exact and heuristic algorithms have been developed to solve this problem (see, e.g., Jordan, 1998). The problem also has qualitative versions: for example, to predict whether the unconditional probability distribution of $X$ is or is not equal to the conditional distribution of $X$ given values for $S$—that is, whether $S$ provides information about $X$—or to predict for any particular value $x$ of $X$ whether its probability increases or decreases when conditioned on a set of values for $S$. Algorithms are available for those problems as well (Pearl, 2000).

*Planning.* It is also possible to use a causal Bayes net to predict the effects of an intervention. We can define an intervention as an action that directly alters or fixes the value of one or more variables in the graph while changing others only through the influence of the directly manipulated variables. That is exactly the kind of prediction we need to make in planning actions to achieve specific goals. There are qualitative versions of these problems as well. General algorithms for these problems are described in Spirtes et al. (1993, 2001) and more accessibly in Pearl (2000). It is possible to compute from the graph alone whether an intervention on one or more variables will change the probability distribution for another variable and to compute the resulting probability distribution. We can sometimes make such predictions even when the available information is incomplete.

Intuitively, these computations can be justified by treating the intervention as a variable with special causal features. For example, consider the canonical case of an intentional human action as an intervention. Take the experimental intervention in our previous example. We intentionally drink wine and then observe the effects on our insomnia. As an intervention, this action has certain distinctive causal features, features that other variables do not have. For example, we believe that our decision to drink directly and exclusively caused us to drink and therefore nothing else did—our intervention fixed the value of the wine-drinking variable, and partying and other variables had no causal effect on that variable. We also believe that our decision to drink only affected other variables, such as insomnia, because it affected drinking itself; it did not, for example, independently increase our insomnia or partying. Moreover, we believe that the decision itself was not caused by other variables in the graph, like wine drinking or partying.

These special causal features of interventions give them a special status in inferring causal structure—that is why experiments are a particularly good way to find out about causal relations. In fact, if an action did not have these features, we could not draw the right causal conclusions. Suppose that, unbeknownst to us, our anxious spouse has replaced the wine in half the bottles in our cellar with a deceptive nonalcoholic grape drink. In this case our intervention to drink does not, by itself, fix whether we actually drink wine, and it does not directly and exclusively cause us to drink or not drink wine; our actual wine drinking is also caused by which bottle we pick out. Our experiment would fail. Or, suppose that doing any sort of experiment makes us so nervous that it keeps us awake—we just cannot take the scientific pressure. We experiment by drinking wine, and sure enough, we stay awake, but we would be wrong to conclude that wine caused our sleeplessness. In this case, the problem is that the intervention affected other variables independently of the variable that was intervened on. Or, suppose the outcome of our experimental intervention subtly influenced our next decision, so that, for example, we were more likely to continue our experiment when the results seemed to encourage us to drink and to curtail it when they did not. That is, downstream variables causally influenced our further interventions. In all three of these cases, we could not draw the right causal conclusions even though we acted.

So, actions that do not have the right causal features should not count as interventions, at least not for purposes of making predictions or uncovering causal relations. Conversely, a variable that did have similar causal features could count as an intervention even if it did not directly involve human action. Although intentional actions are the canonical case of an intervention, from a

formal point of view any variable that has the right causal features can be considered an intervention. These features include the fact that the variable is a direct cause, and the only direct cause, of another variable in the graph, that it fixes the value of that variable, that it is not independently connected to other variables in the graph, and that it is not itself caused by other variables in the graph (see Hausman & Woodward, 1999; Spirtes et al., 1993, Chapter 4). Variables of this special kind might include our own actions, the actions of others, or other events.

The Bayes net formalism provides a way of translating these sorts of intuitions into a formal set of procedures. We expand the graph by adding the new intervention variable and a single new arrow representing the influence of the intervention. The variable that is manipulated is forced to have a fixed value. If we then apply the Markov assumption to the expanded graph, we can predict the effects of the intervention (see Pearl, 2000; Spirtes et al., 1993, 2001).

In most cases, the formal representation of an intervention can be simplified by not explicitly representing the intervention variable. Instead, we specify the value of the manipulated variable that the intervention produces and remove all the other arrows directed in to the manipulated variable (Pearl, 2000, vividly refers to this as "graph surgery"). This simplified representation works because, under the intervention, the manipulated variable has a fixed value; it does not vary as other variables in the system vary. The influence of other variables on the manipulated variable is removed by the intervention, and that is represented by the elimination of the arrows directed into the manipulated variable.

### *Generating Predictions From a Causal Graph: An Example*

To illustrate, we can use the directed acyclic graph in Figure 1 to make two kinds of inferences, one set predicting conditional independence relations among the variables, and the other set predicting the effects of interventions. First, consider the implications of the graph for conditional independence relations. We can expand our earlier probabilistic definition of conditional independence to apply to a set of values of variables. In this way, we can calculate the dependence or independence of two variables in the graph conditional on a set of values of other variables in the graph.

If the causal graph in Figure 1 obeys the Markov assumption, it will encode a variety of conditional independence claims. It implies, for example, that $Z$ and $R$ are independent conditional on the set $\{S\}$ and that $Z$ and $Y$ are independent conditional on the empty set of variables. It does not imply that $Z$ and $Y$ are independent conditional on $\{S\}$. $W$ is independent of all of the other variables conditional on $\{X\}$. $X$ is independent of $Z$ and independent of $R$ conditional on $\{S\}$. Thus, the joint probability distribution represented by the graph in Figure 1 can be written algebraically as the product of the marginal distributions of each variable conditioned on its parents, or $Pr(Z, S, R, Y, W) = Pr(W|X) \times P(X|Y, S) \times Pr(S|Z, Y) \times Pr(R|S) \times Pr(Z) \times Pr(Y)$.

The graph in Figure 1 also represents hypotheses about probabilistic independence if the system were to be subjected to ideal interventions from outside the system. An ideal intervention that fixes a value $X = x$ can be represented in the following way. We introduce a new variable $I$, with two values, one if the intervention takes place, and the other if it does not. We add an arrow going from $I$ to $X$. Then we extend the joint probability distribution on the variables in the graph to include $I$, in the following way. Conditional on $I$ = no intervention, the remaining variables have whatever probability distribution obtains with no intervention. Condition on $I$ = intervention, fix $X$ to $x$ and specify that the probability that $X = x$ is 1. All other variables, except causes of $X$, have their original distribution, but conditioned on $X = x$. Such an intervention fixes a value for one or more of the variables represented in the system. Because the intervention is understood to fix the value of a variable, say $X$, from outside the system, the variables that are represented in the graph as causing $X$ ($S$ and $Y$ in the graph in Figure 1) do not cause $X$ after the intervention.

This fact can also be represented more simply by graph surgery. In this case, we do not represent the intervention variable but instead remove the arrows in the original graph that are directed into $X$ and fix $X = x$. Thus, the causal graph that would result from an intervention that fixes $X$ at the value $x$ is shown in Figure 2.

The probabilities for values of the variables that result from an outside intervention are computed from the conditional probabilities (on $X = x$) associated with the original graph, but the probability of the manipulated variable is changed so that the fixed value has a probability of 1. So, the probability distribution for $S$, $R$, $W$, $Y$, $Z$ that results when $X$ is fixed at $x$ is

$$Pr(Z, S, R, Y, W) = Pr(W|X = x) \times Pr(S|Z, Y) \times Pr(R|S) \times Pr(Z) \times Pr(Y).$$

The manipulation theorem of Spirtes et al. (1993, 2001) says that the conditional independence relations that result from an outside intervention can be determined by applying the Markov assumption to the altered graph as in Figure 2. For example, in the new graph representing the system after the intervention, $S$, $R$, $Y$, and $Z$ are independent of $W$.

Although the formalism may seem complex, the algorithms for computing conditional independence, or for computing the probability of one variable conditional on another, or for computing the probabilities that result from an intervention are efficient. Provided a causal graph is sparse—most pairs of variables are not directly connected by an arrow—computation is very fast. Bayes nets, then, provide a formal and computationally tractable way to generate accurate causal predictions and to design effective causal interventions.

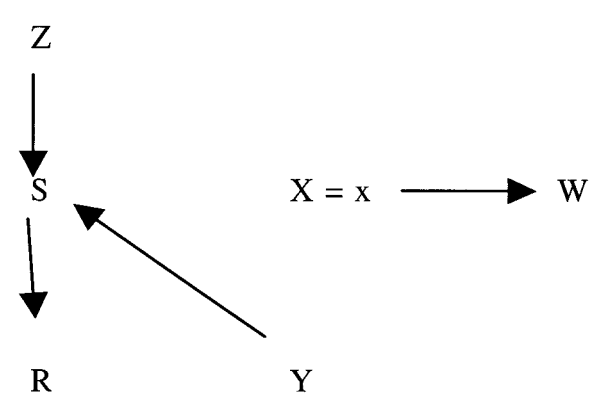


*Figure 2.* The causal graph in Figure 1 with an intervention on $X$.

### *Learning Causal Bayes Nets*

So far, we have shown that causal Bayes nets, that is, directed acyclic graphs with probability distributions that are constrained by the Markov assumption, provide a formalism for representing and using causal relations. Like causal maps, they represent nonegocentric, coherent systems of causal relations—systems that generate accurate predictions and effective interventions. Causal Bayes nets then provide a formal characterization of causal maps. Very recently, some psychologists have suggested that adult causal knowledge might be represented as causal Bayes nets (Glymour & Cheng, 1999; Gopnik, 2000; Gopnik & Glymour, 2002; Lagnado & Sloman, 2002; Sloman & Lagnado, 2002; Rehder & Hastie, 2001; Tenenbaum & Griffiths, 2003; Waldmann & Hagmayer, 2001; Waldmann & Martignon, 1998).

However, the causal Bayes net formalism also suggests ways of representing and understanding how people learn causal knowledge as well as how we use that knowledge. Causal learning is particularly important from the viewpoint of cognitive development. Given a causal graph, we can generate accurate predictions, including predictions about the conditional probabilities of events and about the effects of interventions. This suggests that we could also work backward to generate the graphs from conditional probabilities and interventions. And even in cases where we might not be able to generate the entire graph, we could at least discover aspects of the graphs—for example, we could discover some of the arrows but not others. This would provide us with a method for learning causal Bayes nets from data. To do this, we would have to supplement the Markov assumption with other assumptions.

We describe four general techniques that might be used to learn causal Bayes nets. These include two from computer science (Bayesian and constraint-based learning algorithms) and two from the psychological literature (a causal version of the Rescorla–Wagner rule and the learning rule in Cheng's (1997) causal power theory). The psychological techniques have been applied to causal inference in adults, but they have not been tested in children.

There is a certain trade-off inherent in these two types of techniques. The computer science techniques have generally been applied in data-mining problems, problems that involve information about a wide range of variables, all considered simultaneously. They can infer a very wide range of causal structures from a very wide range of data, but they have psychologically unrealistic memory and processing requirements. The psychological techniques come from the empirical literature on causal learning in adults. They have more realistic memory and processing requirements, but they apply to a much more limited range of causal structures and data.

One important difference between the computational learning methods and current psychological learning methods, in particular, involves the question of determining whether a variable is a cause or an effect. The psychological methods require that the potential causes are discriminated from the potential effects beforehand. Usually this is accomplished with time order information—causes precede effects. But the ordering can also be established by other methods, such as the use of prior knowledge or the knowledge that one variable is being manipulated and the other is not. Then, the psychological learning methods calculate the strength of the causal relations between each potential cause and each potential effect.

The computational learning methods can use this sort of information if it is available, but they can also draw causal conclusions without knowing beforehand which variables are potential causes and which are potential effects. In many cases, they can determine the direction of the causal relation between two simultaneous events. As long as other variables are also measured, these methods can sometimes determine whether $X$ causes $Y$ or $Y$ causes $X$ from the dependencies alone, without relying on time order or prior knowledge. We show later that this fact provides a way of discriminating among these learning methods.

*The faithfulness assumption.* All four of these techniques, and arguably, any technique that could infer Bayes nets from data, must make at least one further assumption, in addition to the Markov assumption itself. This assumption has been formally stated in the context of constraint-based methods, but it is also implicit in other learning methods. It can be stated as follows:

In the joint distribution on the variables in the graph, all conditional independencies are consequences of the Markov assumption applied to the graph.

The principle has been given various names; following Spirtes et al. (1993), we call it the faithfulness assumption. The Markov assumption says that there will be certain conditional independencies if the graph has a particular structure, but it does not say that there will be those conditional independencies if and only if the graph has a particular structure. The faithfulness assumption supplies the other half of the biconditional.

The faithfulness assumption is essentially a simplicity requirement. It might be possible that just by random coincidence, without any causal reason, two causal relations could exactly cancel out each other's influence. For example, going to a party might cause drinking, which causes drowsiness, but the excitement of the party might cause wakefulness, with the result that partying and drowsiness are independent, even though there are causal relations between them—the causal relations cancel one another out. This is a particular example of a phenomenon known as Simpson's paradox in the statistical literature. The faithfulness assumption assumes that such sinister coincidences will not occur.

A causal learner in a Simpson's paradox situation is like someone looking into a Viewmaster. Three-dimensional objects produce particular patterns of two-dimensional images at each eye. The Viewmaster works by presenting each eye with the image that would have been produced by a three-dimensional object, with the result that the viewer sees an object in depth. The visual system makes a kind of faithfulness assumption; it assumes that the observed visual relations were produced by a three-dimensional structure even though, in fact, they were not.

It has been shown that for absolutely continuous probability measures on the values of linear coefficients of linear models, faithfulness holds with a probability of 1, and similarly for absolutely continuous probability measures on the conditional probabilities (of each variable on each vector of values of its parents) in models with discrete variables (Meek, 1995; Spirtes et al., 1993). It is easy to construct violations of the faithfulness assumption mathematically. However, in nondeterministic or noisy systems, we would be extremely unlikely to encounter a set of events that violated the assumption, just as we would be extremely unlikely to encounter a phenomenon like the Viewmaster in the natural world.

*The search problem.* The Markov assumption and the faithfulness assumption are like the geometric and optical assumptions

that allow the visual system to solve the spatial inverse problem. By making these two quite general assumptions about the causal structure of the world and the relation between causation and conditional independence, we can provably solve the causal inverse problem for a great variety of types of causal structure and types of data.

This solution relies on the fact that, according to the Markov and faithfulness assumptions, only some causal graphs and not others are compatible with a particular set of conditional probabilities of particular variables. These assumptions constrain the possibilities; they say whether a particular graph is or is not consistent with the data. This leaves, however, two further problems. There is the algorithmic problem of finding a way to efficiently search through all the possible graphs and discard those that are inconsistent with the data. There is also the statistical problem of estimating the probabilities from the data.

Again, we can draw the analogy to vision. Given certain assumptions about geometry and optics, mathematical theories can say, at least in part, which representations of objects are consistent with a particular set of perceptual data. Marr (1982) called this the computational level of representation. For example, the geometrical assumptions in mathematical vision science say that a three-dimensional object that is projected onto 2 two-dimensional retinas will lead only to certain disparities between those images and not others. In practice, however, computer vision systems (or, for that matter, biological vision systems) must also find procedures that allow them to compute the object representations from the retinal data in a reasonably efficient way. Marr (1982) called this the algorithmic level of representation. Vision scientists use the geometrical theory that relates depth to disparity to help design search procedures in computer vision programs and to help discover the search procedures that are actually used by the human visual system.

In the causal case, one way of making the search tractable is to limit the possibilities by using other kinds of information or assumptions about the graphs. For example, temporal order information can rule out certain possibilities; $Y \rightarrow X$ should not appear in the graph if $X$ always comes before $Y$. Similarly, other kinds of prior knowledge can influence the search. We may already know, for example, that $X$ causes $Y$ or that $X$ does not cause $Y$, and that means that we must include or exclude an arrow from $X$ to $Y$ in the graphs. Or someone else may explicitly tell us that $X$ causes $Y$ or give us other facts about the graphs. Similarly, we may know from other sources that the graphs have specific types of structures or specific parameterizations and so restrict our search appropriately.

However, in the computer science literature, efficient search procedures have been developed that make minimal additional assumptions, although they can incorporate other types of knowledge if they are available. The procedures can be applied to purely observational data, to data that include interventions, to combinations of the two sorts of data, to continuous variables, to discrete variables, and to certain combinations of discrete and continuous variables, with and without a range of prior knowledge. These learning procedures include Bayesian methods (Heckerman, Meek, & Cooper, 1999) constraint-based methods (Scheines, Spirtes, Glymour, & Meek, 1994), and various combinations of the two.

*Computational approaches: Bayesian methods.* In general, Bayesian causal learning methods have the same structure as Bayesian methods in statistics. The possible causal hypotheses, represented by Bayes nets, are assigned a prior probability. This probability is then updated, given the actual data, by the application of Bayes theorem. Typically, we accept or conjecture the hypothesis with the highest posterior probability, but we will also know the probability of other hypotheses. In principle we can, if we choose, sum the probability of any particular causal connection over all of the hypotheses.

In detail and ideally, a prior probability measure is imposed on every directed acyclic graph of interest. A family of possible joint probability distributions for the variables is assumed, as a function of a finite set of parameters associated with each graph. For example, if a distribution family is assumed to be Gaussian, each variable is assumed to be a linear function of the values of its parents plus a normally distributed error term. The parameters are then the linear coefficients and the variance of each variable. If the variables are all discrete, the joint distribution is typically assumed to be multinomial, and the parameters are the probabilities of each variable conditional on each vector of values of its parents. The directed acyclic graph together with a complete set of parameter values determine a unique joint probability distribution on all of the variables. This probability distribution in turn determines a sampling distribution, assuming the Bayes net describes an independent probability distribution for each unit in the sample.

Putting the pieces together, when integrated, the prior probability distribution over the graphs, multiplied by the probability distribution over the parameters conditional on each graph, multiplied by the sample probability conditional on each graph and each possible set of parameter values for that graph, results in a prior probability distribution for the sample. Under various technical assumptions, the sampling distribution conditional on any given graph can be quickly computed (Heckerman, 1995).

Ideally, Bayes theorem is then applied to compute the posterior probability distribution over the graphs, conditional on the data. In practice, because the number of directed acyclic graphs grows superexponentially with the number of variables, heuristic greedy algorithms are used instead. One starts with an arbitrary graph, computes its posterior probability, computes the posterior probabilities of a specified set of alterations of the initial graph (adding, deleting, or reversing arrows), chooses the alteration with the highest posterior probability, and repeats the process until no more improvements are found.

*Constraint-based methods.* Constraint-based methods work quite differently. In these methods, the dependence or independence between each set of variables is calculated from the data, as the algorithms require them. These dependence relations are de-

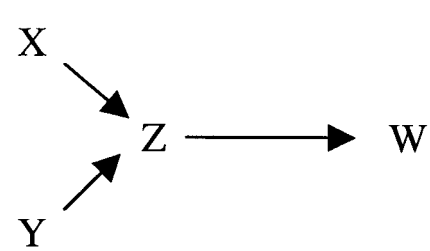


*Figure 3.* Discovering a graph by constraint-based methods: The true structure.

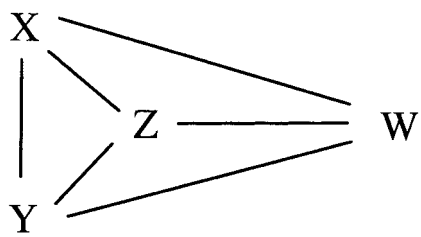


*Figure 4.* Discovering a graph by constraint-based methods: Step 1.

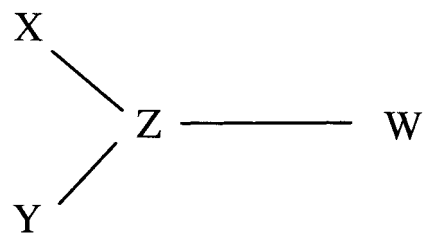


*Figure 6.* Discovering a graph by constraint-based methods: Step 3.

termined by standard statistical tests of significance applied to the actual data. Graphs are constructed that are consistent with those dependence and independence relations, step by step. The TETRAD algorithms (Scheines et al., 1994) are typical of constraint-based discovery procedures. They are most clearly described by an example. Suppose the unknown structure to be discovered is as in Figure 3.

Note that according to the Markov and faithfulness assumptions, this graph implies that the following independence relations, and only these, hold: $X \perp Y$; $W \perp \{X, Y\}|Z$. We are given data on $X$, $Y$, $Z$, and $W$ for a sample of units drawn from an unknown probability distribution and make the Markov assumption and faithfulness assumption about the graph in Figure 3. We are also given the information that the probability distribution belongs to a family of probability distributions—say normal or multinomial—but no other information. In particular, there is no information about time order. We cannot recover the entire graph, but we can discover the following: either $X$ causes $Z$ or there is an unmeasured common cause of $X$ and $Z$. This is similar for $Y$ and $Z$. $Z$ causes $W$. Further, we can discover that there is no unmeasured common cause of $Z$ and $W$. Here is how we are able to learn this.

1. Form the complete, undirected graph on all of the variables, as in Figure 4.

2. Test each pair of variables for independence. Eliminate the edges between any pair of variables found to be independent. The result is Figure 5.

3. For each pair $U$, $V$, of variables connected by an edge, and for each variable $T$ connected by an edge to one or both of $U$, $V$, test whether $U \perp V|T$. If an independence is found, remove the edge between $U$ and $V$. The result is Figure 6.

4. For each pair $U$, $V$ of variables connected by an edge and each pair $T$, $S$ of variables, each of which is connected by an edge to either $U$ or $V$, test the hypothesis that $U \perp V|\{T, S\}$. If an independence is found, remove the edge between $U$, $V$. In the graph of Figure 5, $Z$ is adjacent to $W$ and $Z$ has two adjacent variables, but $Z$ and $W$ are not independent conditional on $\{X, Y,\}$ and no change is made. This part of the procedure stops.

5. For each triple of variables $T$, $V$, $R$ such that $T - V - R$ and there is no edge between $T$ and $R$, orient as $T\text{o} \rightarrow V \leftarrow \text{o}R$ if and only if $V$ was not conditioned on when removing the $T - R$ edge. (This results in Figure 7.)

6. For each triple of variables, $T$, $V$, $R$ such that $T$ has an edge with an arrowhead directed into $V$ and $V - R$, and $T$ has no edge connecting it to $R$, orient $V - R$ as $V \rightarrow R$. The final result is Figure 8.

The o marks indicate that the procedure cannot determine whether the association is produced by a causal arrow from one variable to the other, as from $X$ to $Z$, or by a common unobserved cause of $X$ and $Z$, or both.

The general algorithms for constraint learning of causal Bayes nets are given in Spirtes et al. (1993, 2001), along with proofs of their asymptotic correctness under the Markov and faithfulness assumptions. Given these assumptions and given sufficient data, these algorithms will almost certainly come to the correct conclusion about which possible causal structures are consistent with the data.

Constraint-based search methods can use prior knowledge about the existence or absence of particular causal connections. For example, if a constraint-based program such as TETRAD II (Scheines et al., 1994) is told that $Z$ occurs later than $X$ and $Y$ occurs earlier than $X$, then in testing whether there is a direct connection $Y \rightarrow X$, the program will not test their independence conditional on $Z$. The program can also use prior but uncertain knowledge about causal connections. For example, if the program

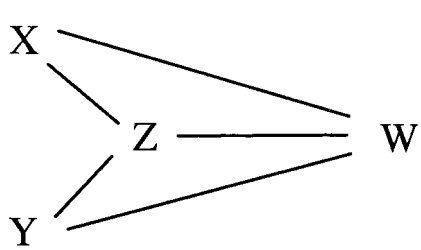


*Figure 5.* Discovering a graph by constraint-based methods: Step 2.

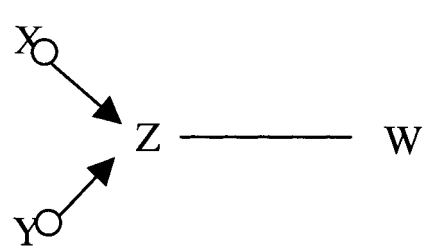


*Figure 7.* Discovering a graph by constraint-based methods: Step 4.

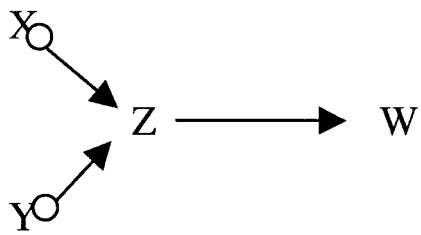


*Figure 8.* Discovering a graph by constraint-based methods: Step 5.

is told that there is some chance that $Y$ may directly cause $X$, it can adjust the significance level in statistical decisions as to whether $Y$ and $X$ are independent or independent conditional on other variables. These uncertain prior degrees of belief are handled more elegantly in Bayesian search methods.

*Psychological approaches: The causal Rescorla–Wagner method.* Work in the psychological literature also suggests methods for learning causal structure from data. Associative learning, for example, has been proposed as a method of learning causal relations (Shanks & Dickinson, 1987). It is important to distinguish between this causal interpretation of associative learning and associative learning per se. Classical associative learning theories assume that organisms simply make associations without learning about anything independent of the associations themselves. Such theories have classically been applied to learning in animals, and they have also been extensively applied to learning in children (see, e.g., Elman et al., 1996; Thelen & Smith, 1996). Instead, some investigators have recently proposed that human adults use associative learning rules to infer underlying causal relations. One could think of these accounts as techniques used to infer a causal graph from data.

The basic principle underlying such an account is that the strength of the association between two variables, calculated by associative learning rules, indicates the probabilistic strength of the causal connection between them. Presumably, we could construct a causal graph by combining information about pairs of variables. This information could then be combined with other types of causal information. Finally, it could be translated into a set of instructions for intervening on one event to bring about another. However, these links among association, causal structure, prior knowledge, and intervention have not been made explicit in this literature.

The most influential associative learning procedure is due to Rescorla and Wagner (1972). The Rescorla–Wagner (RW) procedure estimates that the associative strength of potential cause $C_i$ with the effect, $E$, after trial $t$ is $V_i^{t+1} = V_i^t + \Delta V_i^t$, where $\Delta V_i^t$ is given by

$$\Delta V_i^t = \begin{cases} 0, \text{ if the cause, } C_i\text{, does not appear in case } t. \\ \alpha_i\beta_1\left(\lambda - \sum\limits_{\text{Cause } C_j \text{ appears in case } t} V_j^t\right), \\ \qquad \text{if both } C_i \text{ and } E \text{ appear in case } t. \\ \alpha_i\beta_2\left(0 - \sum\limits_{\text{Cause } C_j \text{ appears in case } t} V_j^t\right), \\ \qquad \text{if } C_i \text{ appears and } E \text{ does not in case } t. \end{cases}$$

Unlike constraint-based algorithms, the RW algorithm gives a trajectory of associations or estimates of causal strength as the data are acquired step by step. The RW process is often compared with other learning theories through its long run behavior, or equilibria (Cheng, 1997; Danks, 2003). A vector of associative strengths $\mathbf{V} = < V_0, \ldots, V_n >$ (one dimension for each cause) is an equilibrium of the RW model for a probability distribution if and only if $\forall i[E(\Delta V_i) = 0]$. That is, a strength vector is an equilibrium if and only if, for every cause, the expected value of the change in the associative strength of that cause with the outcome is zero. Cheng (1997) characterized a great many cases in which the equilibria of the RW procedure learning the effect of $X$ and $Y$ on $Z$ are

$$Pr(Z|X, \sim Y) - Pr(Z|\sim X, \sim Y)$$

and

$$Pr(Z|\sim X, Y) - Pr(Z|\sim X, \sim Y).$$

Danks (2003) gave a fully general characterization of the equilibria.

When the potential causes occur before the effect and there are no unmeasured common causes of the potential causes and the effect, the associative strengths might be interpreted as estimates of the strengths of causal connections in a Bayes net. In fact, thinking about the RW rule in the context of Bayes nets gives an interesting evolutionary and normative explanation for the very existence of the rule. RW works because, in many cases, it recovers the correct causal structure of the world.

Even in such special cases, however, the interpretation is problematic for a variety of reasons. Cheng (1997), for example, showed that adult subjects are sensitive to ceiling effects not captured by equilibria of the RW procedure. Further, the learning model can only account for interactive causes by treating the combination of interacting causes as a separate cause.

*Psychological approaches: Cheng's (1997) power PC method.* Cheng (1997) has recently proposed a theory of causal representation and learning for adult humans, the power PC theory. We consider the representation and the learning theory separately. Although it was empirically motivated and developed independently, the power PC theory representation is equivalent to a Bayes net representation with a special parametric form for the probabilities. Networks so parameterized are known as *noisy-or gates* and *noisy-and gates* in the computer science literature, and they can be chained together into any directed acyclic graph and can include unobserved common causes (Glymour, 2001). The Markov condition necessarily holds if noisy-or and/or noisy-and gates are chained together into a directed acyclic graph. Novick and Cheng (in press) have given a much more intricate set of parameterizations of causal models with interaction, without precedent in the computer science or statistical literatures, and these structures, too, can be chained together in networks. Cheng's theory, then, proposes that adult human causal representation involves a particular type of causal Bayes net with a particular parameterization.

Cheng (1997) also provided a method for learning such graphs, and the causal power of the arrows, from data. She provided a direct estimate of causal strength, again assuming that the potential causes are discriminated from the effects, that there are no unobserved common causes of the observed potential causes and the effect, and that the potential causes do not interact to produce the effect. This estimate differs from the estimate provided by the RW rule. Cheng's estimator for a generative cause $A$—one whose presence increases the probability of the effect, $E$—is

$$\frac{fr_F(E|A) - fr_F(E|\sim A)}{[1 - fr_F(E|\sim A)]}.$$

The frequency $fr_F$ is for a focal set of cases in which subjects judge $A$ to be independent in probability of any other causes of $E$. This focal set is defined psychologically—it is the set of cases in which the subject believes that the potential cause being assessed is independent of other potential causes of the effect. It is not necessarily derived from any particular objective features of the data, though, of course, one assumes that the data affect the subject's beliefs. Cheng (1997) observed, in particular, that when the values of $A$ are the result of interventions, subjects tend to regard $A$ as independent of other causes of $E$. A different estimator is given when the cause is preventive and lowers the probability of the effect.

Made into a learning rule, this estimator is asymptotically correct for the parameterization of Bayes nets of the kind Cheng (1997) specified—in particular, in cases in which there are no unobserved common causes of potential causes ($A$ above) and effects ($E$ above). (This is assuming that, in the focal set, the potential causes really are independent in probability from other potential causes of the effect.) Her estimator of the efficacy of $A$ in these cases equals, asymptotically, the probability that the effect occurs conditional on $A$ occurring and no other cause of the effect occurring. Novick and Cheng (in press) gave related estimation rules for interactive causes of a variety of types. Generalizations of Cheng's rule have been shown to be able to correctly estimate her generative causal power parameters in certain cases in which the potential cause $A$ and the effect $E$ are both influenced by an unobserved common cause (Glymour, 2001).

*The statistical problem: Fictional sample sizes and learning rates.* We mentioned above that, in addition to the search problem, learning a causal Bayes net also presents a statistical problem. Before we can infer causal structure from conditional probabilities, we need to be able to infer those conditional probabilities from data about frequencies. In the experiments we subsequently describe, the number of trials is typically very small, so that, from a statistical point of view, the data provide little information about underlying probabilities. Nonetheless, as we show, children are very willing to make causal inferences from such small samples.

At least three of these four learning algorithms must make some further qualitative assumption to account for these judgments. For Bayesian learning algorithms, the further assumption is an informative prior probability distribution over the graphs and parameters—that is, a distribution that does not give uniform prior probabilities to all hypotheses. In other words, children might use their prior knowledge that some graphs are more likely than others to help infer probability judgments from the limited new data (Tenenbaum & Griffiths, 2003). An informative prior probability is equivalent to assuming the prior probabilities used were obtained by applying Bayes' rule to an uninformative prior probability and a fictional sample. Constraint-based procedures must assume that each of the observed cases is multiplied by the same number to form a fictive sample size—for example, each case is treated in statistical inference as if it were 100 cases. In other words, children might assume that the small samples they see are representative of the actual distribution, so, for example, if they see that A is associated with B on 1 trial, they assume that they would see the same thing on 100 trials. Similarly, the RW model must set one of its parameters—the learning rate—to a very high value.

Cheng's (1997) learning method does not use a statistical procedure or a learning rate but estimates causal powers directly from the observed frequencies. This has the advantage that no extra assumption is needed to address small sample cases. However, it does not provide a dynamical account of how causal inference might improve as children gain a larger data set. Danks (Danks, Tennenbaum, & Griffiths, 2003) has recently shown that there is a dynamical learning procedure, analogous to the RW updating rule, that converges to Cheng's generative causal powers. This procedure supplies a learning rate analogous to the RW parameter.

*Learning Bayes nets with unobserved variables.* So far, we have been considering learning models that recover Bayes nets in which all the variables are observed. Bayes net representations can also involve unobserved variables, and there are procedures for learning about such unobserved variables from the data. These procedures are particularly interesting from a psychological point of view. The theory theory, for example, proposes that children can infer unobserved variables, such as internal mental states, from patterns of data. How might such learning be possible?

The Markov assumption is not assumed to hold for observed variables alone, and in fact, it will not hold for just the observed variables when there are unobserved common causes, that is, when there are unobserved variables that directly influence two or more observed variables. Consider, for example, the structure shown in Figure 9, where $U$ is an unobserved variable. The Markov assumption, applied to the graph in Figure 9, implies $X \perp \{Z, R\}$ and $R \perp \{X, Y\}$ and no other independence or conditional independence relations among $\{X, Y, Z, R\}$ alone.

There is no directed acyclic graph on the observed variables alone that implies these, and only these, independence relations. It is not hard to see why. $Y$ and $Z$ are dependent, so either $Y \rightarrow Z$ or $Z \rightarrow Y$. This must be true because neither $X$ nor $R$ can be common causes of $Y$ and $Z$, as $X$ is independent of $Z$ and $R$ is independent of $Y$. Suppose $Y \rightarrow Z$. Then, because $X$ and $Y$ are dependent, either $X \rightarrow Y$ or $Y \rightarrow X$, or else one or both of $R$, $Z$ are common causes of $Y$ and $X$. $R$ cannot be a cause of $Y$ because $R$ and $Y$ are independent, and $Z$ cannot be a cause of $Y$ because, by the supposition, $Y$ causes $Z$, and the graph would be cyclic. So, either $X \rightarrow Y$ or $Y \rightarrow X$. However, in either case, by supposition, $Y \rightarrow Z$, and therefore, by the faithfulness and Markov assumptions, $X$ and $Z$ are not independent. However, this contradicts the first of the independence relations above: $X$ and $Z$ are independent. The alternative supposition, $Z \rightarrow Y$, leads to the same conclusion by a parallel argument. In causal contexts, the Markov and faithfulness assumptions apply only to *causally sufficient* sets of variables—that is, to sets of variables that include every common cause of their members. This feature of causal Bayes net representations turns out to be useful in learning.

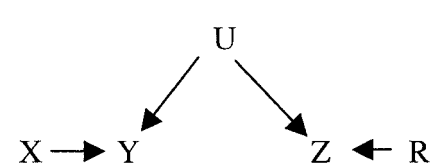


*Figure 9.* A graph with an unobserved variable.

Unobserved common causes produce extra dependencies among variables. Conversely, violations of the faithfulness assumption (which can occur in deterministic systems without noise) produce extra independencies (or conditional independencies) among variables. We can algorithmically decide whether the independencies and conditional independencies violate faithfulness, and if faithfulness is not violated, we can determine whether the dependencies violate the Markov assumption. If they do, we can conclude that there must be an unobserved common cause. In short, if there are dependencies in the data that could not be generated by a Bayes net involving the observed variables, we conclude that there must be some unobserved common cause or causes that is responsible for the additional dependencies.

Consider once more the example in Figure 9 with the independence relations: $X \perp \{Z, R\}$ and $R \perp \{X, Y\}$. If we apply the procedure previously illustrated for constraint-based learning, we first form the complete undirected graph, then remove edges between $X$ and $R$, $Y$ and $R$, $Z$ and $X$, because each of these pairs is independent. The result is the undirected graph: $X - Y - Z - R$.

There are two triples whose terminal variables are not connected by an edge: $X - Y - Z$ and $Y - Z - R$. In removing the $X - Z$ edge we did not condition on $Y$, and in removing the $Y - R$ edge we did not condition on Z, so these triples are oriented: $X\text{o} \rightarrow Y \leftarrow \text{o}Z$ and $Y\text{o} \rightarrow Z \leftarrow \text{o}R$. Hence, the final result of the procedure is $X\text{o} \rightarrow Y \leftrightarrow Z \leftarrow \text{o}R$. The double-headed arrow indicates that $Y$ and $Z$ must be connected by a common cause that is neither $X$ nor $R$.

Constraint-based procedures, then, can sometimes identify the presence of unobserved common causes. At present, Bayesian methods are much more limited in this respect. If you start out with specific alternative graphs that involve unobserved variables, Bayesian methods can use the data to compare the posterior probabilities of those graphs with the posterior probabilities of other graphs. However, there are as yet no Bayesian methods that do a general search for the presence of unobserved variables.

## Causal Bayes Nets as a Psychological Model of Children's Learning

The concept of causality is an intricate web relating ideas about association, independence, conditional independence, and intervention. The causal Bayes net representation may not capture the whole of this web of ideas, but it captures a great deal of it. Our psychological hypothesis is that people, and children in particular, represent causal relationships in ways that can be described as causal Bayes nets. Moreover, they apply learning procedures to construct new causal representations from observations of correlations and interventions. This hypothesis provides a framework for more specific explanations of specific causal learning abilities. It can help explain how children can derive correct predictions of the effects of interventions from passive observation, how they can integrate correlational data and data from interventions, and how they can combine prior knowledge and new observations to discover new causal relations. It suggests processes through which children can correctly postulate unobserved variables. No other representational proposal that we know of allows for this range of inferences and predictions.

Of course, we are not proposing that children have conscious knowledge of these representations or learning procedures or that they have little graphs in their heads. Our proposal is that causal knowledge and learning can be represented in this way at the computational level (Marr, 1982). We can say that children use Bayes nets or infer Bayes nets from data in much the same way we can say that the visual system uses representations of three-dimensional structure or infers these representations from stereoscopic disparities. This leaves open the further question of how these computations are actually implemented in the brain (what Marr, 1982, called the implementation level). In the case of vision, at least low-level vision, we have some ideas about how neural connections actually implement these computations. We hope that, at some future point, a similar project will be fruitful in the causal domain. However, as in vision, such a project depends on specifying the representations and computations first.

Our hypothesis can be differentiated from other hypotheses about causal learning in children. It is possible that young children do not learn genuinely new causal relations but initially rely on a few innate domain-specific causal schemas that are later enriched (see, e.g., Atran, 1990; Leslie & Roth, 1993; Spelke et al., 1992). A related, though somewhat richer, view is that children only (or primarily) use substantive assumptions, such as spatial contact and temporal order, to infer new causal relations and do not learn about new causal relations from information about correlation (Ahn et al., 2000). Alternatively, it is possible that, even if young children do learn, learning is restricted to the mechanisms of classical or operant conditioning. Again, a related but somewhat richer hypothesis is that young children simply associate events but do not infer genuine causal relations between them (see, e.g., Elman et al., 1996; Thelen & Smith, 1994). Alternatively, children might only use trial and error or imitative learning to determine the direct causal consequences of their own actions or those of others.

None of these hypotheses accounts for, or allows, most of the features of causal inference described earlier. In fact, even if children used all these methods of learning together, the inferences they made would still be restricted in important ways. We present new empirical evidence that demonstrates that very young children can, in fact, make causal inferences that require more powerful learning mechanisms, such as all four of the formal learning mechanisms we described in the previous section.

This general hypothesis is nonetheless consistent with a variety of other specific hypotheses about how causal relations are learned, including the four different formal learning models we described above and many variations and extensions of those models. We further hypothesize, more specifically, that the causal learning mechanisms that are involved in children's cognitive development lie somewhere between those proposed by the computational and psychological types of learning methods. That is, they are more powerful and general than the psychological learning mechanisms that have currently been proposed, but they are, at least in some respects, more constrained than the normative computational learning mechanisms.

We make this hypothesis because children do, in fact, seem to be able to learn more complex causal structures from a wider variety of data than the current psychological models address. For example, the literature on everyday psychology suggests that children learn complex causal maps relating beliefs, desires, emotions, and perceptions (e.g., Gopnik & Wellman, 1994). This literature also suggests that children might make causal inferences without discriminating potential causes and potential effects beforehand. In many psychological cases, mental states can be both causes and

effects, and it is not obvious whether one person's actions caused another's or vice versa. Moreover, mental states are not usually directly observed.

However, it is also eminently possible that children are not capable of the same types of causal learning as experienced and educated adults. Before the experiments we describe here, there was no evidence that children were even able to use the same formal causal learning mechanisms demonstrated in human adults, let alone more general mechanisms. If children do not have such mechanisms available, we have to find some other explanation for the development of the causal knowledge encoded in everyday theories. Perhaps, contrary to our hypotheses, this knowledge is innate rather than learned.

### *Testing Strategies*

Before we describe our experiments, however, we clarify how they are related to the Bayes net formalism. The formalism is not itself a psychological hypothesis. Instead, it is a normative mathematical account of how accurate causal inference is possible, whether this inference is performed by children, computers, undergraduates, or sophisticated adult scientists. Again, we return to the vision analogy. The relation between depth and stereo disparity is not a psychological fact but a geometrical and optical fact, one that allows accurate inferences to be made by any kind of visual system, human, animal, or artificial.

The vision science formalism also involves several interwoven representations and assumptions. The mathematical account of stereo includes a geometrical representation of the objects we see and a set of geometrical and optical assumptions about how those objects produce patterns of visual disparity. It also includes the equivalent of the faithfulness assumption—that is, the assumption that the disparities we see were actually produced by the geometry of the object and its geometrical and optical relations to two-dimensional images. These three aspects of the formalism allow us to create a variety of algorithms that let us work backward from disparity to depth in a reasonably efficient way.

The Bayes net representation has three very similar essential pieces—the graphical representation of causal relations, the Markov assumption connecting each graph with constraints on probability distributions, and the faithfulness assumption, which assumes that the probabilities are, in fact, produced by the representations and the Markov assumption. Those essential and interrelated pieces can be used to create specific learning algorithms. The role of the Markov condition in causal reasoning can scarcely be tested in isolation from other assumptions, any more than the geometric assumptions of optics can be tested apart from a three-dimensional geometric representation of objects.

In vision, we assume that the representations and assumptions are unconscious—our subjects would hardly be informative if we asked them explicitly whether three-dimensional objects are consistent with horizontal and vertical disparities. We also do not test different parts of the geometry separately. Instead, in psychophysics, we test the entire model of depth from disparity by getting people to look at different patterns of disparities and determining whether they see depth. Such evidence tells us whether the visual system uses the geometrical and optical information, and it can help us decide which particular search algorithms might be used.

We make the same assumptions about causal learning. We obviously cannot ask 3-year-olds explicitly whether they think that particular causal structures are consistent with particular conditional probabilities. Moreover, we are almost certain that adults would make mistakes if they were asked to make explicit judgments of this kind, even if they unconsciously used the correct assumptions in their implicit causal judgments. Instead, the experiments with children that we describe here are a kind of cognitive psychophysics. We present children with various patterns of evidence about conditional probabilities and see what causal conclusions they draw. We can then see how much, or how little, these unconscious inferential procedures are in accord with the normative mathematical account, again just as in the case of vision, and we can discriminate among particular learning procedures.

In our first study, we show that children as young as 30 months can implicitly use conditional dependence information to learn a causal map. In the second study, we offer further evidence for this ability from a paradigm—backward blocking—that directly contradicts the causal RW model. Then we show that children can make normatively accurate inferences to solve a problem that is outside the scope of Cheng's (1997) published theory and causal RW models but can be handled by constraint-based and Bayesian methods. This problem involves determining the causal relation of two simultaneous events—did $X$ cause $Y$, or did $Y$ cause $X$? Then we extend this paradigm to provide further evidence that directly contradicts the causal RW model. Finally, we describe preliminary evidence that children can also solve another problem that is outside the scope of published Cheng methods and RW methods but can be handled by other Bayes net learning methods. This problem involves inferring the existence of an unobserved common cause.

### *Causal Learning in Young Children: Experimental Evidence*

How could we test these ideas empirically? In particular, what could we do to discover how, or even whether, children use information about conditional dependence to construct accurate causal maps? We need methods that allow us to expose children to patterns of evidence in a controlled way and to see whether they draw genuinely novel and genuinely causal conclusions on the basis of that evidence. We need a kind of developmental cognitive psychophysics. We describe experiments that use two such methods: the blicket detector and the puppet machine.

### *Inferring Causal Maps From Conditional Dependence: The Blicket Detector*

Gopnik and colleagues (Gopnik & Esterly, 1999; Gopnik & Nazzi, 2003; Gopnik & Sobel, 2000; Nazzi & Gopnik, 2000) devised the "blicket detector" to explore children's causal learning. The blicket detector is a square wooden and plastic box that lights up and plays music when certain objects, but not others, are placed on it. (In fact, the machine is secretly controlled by a human confederate, but neither adults nor children guess this.) This apparatus appears to present children with a new, nonobvious, causal relation. Some objects (which we call *blickets*) have the causal power to make the machine go, and some do not. We can then expose children to different patterns of evidence about the blocks

and the detector and discover what causal inferences they draw. A first set of studies (Gopnik & Nazzi, 2003; Gopnik & Sobel, 2000; Nazzi & Gopnik, 2000) demonstrated that young children could, in fact, learn about this new causal relation and use that knowledge to categorize the objects.

In the next set of experiments, we explored whether children could use evidence about conditional dependence and independence to infer these new causal relations. Gopnik, Sobel, Schulz, and Glymour (2001) presented 3- and 4-year-old children with the blicket detector after a familiarization period in which the experimenter told children that the machine was "a blicket machine" and that "blickets make the machine go." Children were then presented with two types of tasks. In the "one-cause" tasks, the experimenter first put Object A on the detector, which activated it. Then he put Object B on the detector, which did not activate it. Then he placed both objects on the machine simultaneously twice in a row. The machine activated both times. Finally, he asked the children whether each object, individually, was a blicket or not (see Figure 10).

Children might have solved the one-cause task by simply picking the object that made the machine go off more frequently. To control for this possibility, we included a "two-cause" task: The experimenter placed Object A on the machine by itself three times. Each time, Object A activated the machine. Then the experimenter placed Object B on the machine by itself three times. It did not activate the machine the first time, but did activate it the next two times. Again, children were asked whether each object individually was a blicket. In this experiment, as in the one-cause task, Object A set the machine off all three times and Object B set it off two out of three times. If children responded differently in the two tasks, they could not simply be responding to frequencies. Moreover, this task also controlled for the possibility that children were using other simple perceptual strategies, such as picking the object that activated the machine first.

In the one-cause task, children said that Object A was a blicket significantly more often than they said that B was a blicket (96% vs. 41%). In contrast, in the two-cause task they were equally likely to say that A and B were blickets (97% and 81.5%). Moreover, they said that Object B was a blicket significantly more often in the two-cause task than in the one-cause task.

Notice, however, that some children in the one-cause task did say that both objects were blickets. This may have been the result of a general tendency to say "yes," which is common in very young children. In a second experiment, 30-month-olds were given a slightly modified procedure. Children were asked to make a forced choice between the two alternatives: "Which one is the blicket, A or B?" The results were similar; children chose Object A significantly more often than Object B in the one-cause task (78% vs. 22%), they chose each object equally often in the two-cause task (47% vs. 53%), and they chose Object B significantly more often in the two-cause task than in the one-cause task.

The experimental condition—that is, the one-cause task—can be analyzed in several ways. Let $A$ and $B$ be binary variables representing the location of Objects A and B (present or absent on the detector), and let $D$ be a binary variable for the state of the detector (on or off). We have that the frequency ($fr(A, D) \neq$

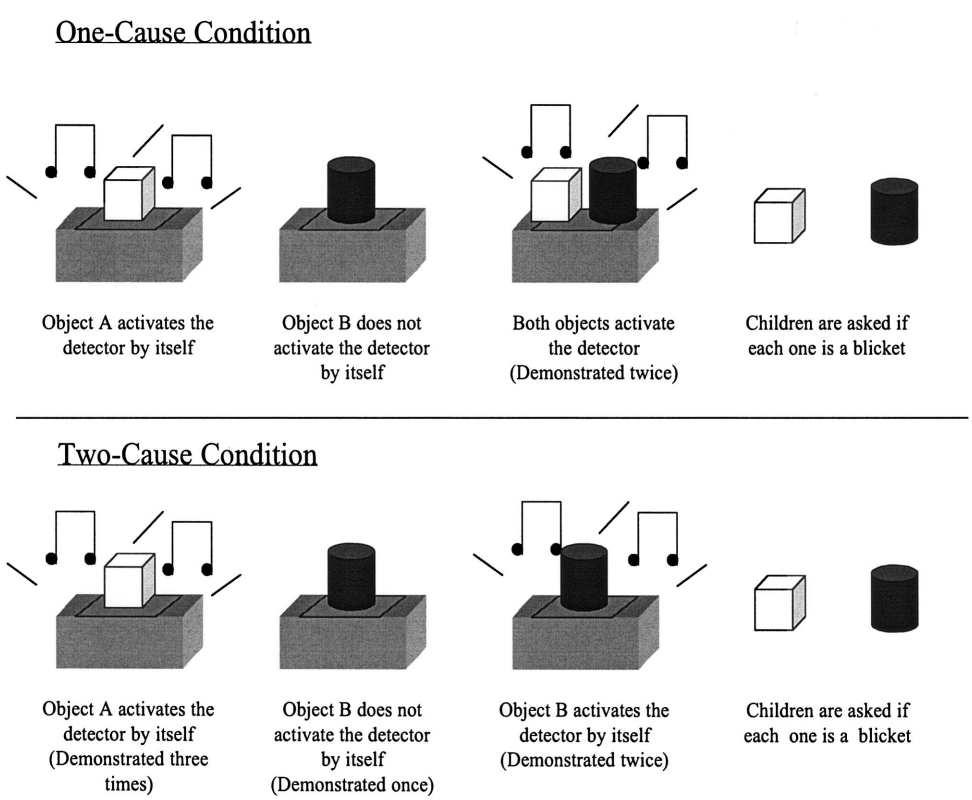


*Figure 10.* Procedure used in Gopnik, Sobel, Schulz, and Glymour (2001, Experiment 1).

$fr(A)fr(D)$ and $fr(B, D) \neq fr(B)fr(D)$. That is, $A$ and $B$ are each dependent in probability on the state of the detector. Further, $fr(A, B) \neq fr(A)fr(B)$—that is, $A$ and $B$ are not independent. Also, $fr(B, D|A) = fr(B|A)fr(D|A)$. The state of the blicket detector is independent of the presence and of the absence of Object B conditional on the presence of Object A. Finally, $fr(A, D|B) \neq fr(A|B)fr(D|B)$. Conditioning on the value of $B$ does not make $A$ and $D$ independent.

Applied to this case, the constraint-based learning algorithm for Bayes nets constructs the following model (shown in Figure 11), (provided it is assumed that the interventions eliminate the possibility of common causes of Object A and the detector, and of Object B and the detector, and the sample size is given a large fictitious multiplier). This graph represents all the possible causal graphs that are consistent with the Markov and faithfulness assumptions and the patterns of conditional dependence we just described.

The graph says that $A$ causes $D$ and $B$ does not. (It also says that there is some undetermined, and perhaps unobserved, causal link between $A$ and $B$, represented by the circles and the ends of the edge connecting those variables. In fact, there is such a link, namely the experimenter, who usually puts both of the blocks on the machine at the same time.) The experiment can also be explained on a number of other hypotheses. The power PC estimate of generative causal power is 1 for $A$ and 0 for $B$. With a very high learning rate (analogous to the sample size multiplier in the constraint search), RW yields a high association of $A$ with $D$ and a low association of $B$ with $D$.

*Designing a new intervention.* In these experiments, we assumed that children understood our original instructions about the causal power of the blocks and were indeed using the word *blicket* to identify the blocks that made the machine go. They were constructing a causal map of the blocks and the detector by using one of the methods we described above, including perhaps the causal version of RW. However, it was also possible that children were not making a causal inference at all, even a causal inference using RW, but were simply associating the word *blicket* with the effect and further associating the block with the effect. They need not have understood that the blicket actually made the machine go. How could we ensure that children really were reasoning causally and creating a causal map, rather than simply associating the word *blicket* and the effect?

Two things could distinguish a genuinely causal map from a simple association. First, causal reasoning implies a set of predictions about interventions. If $A$ is causally related to $B$, then taking action to influence $A$ should influence $B$. If children really think that the blickets make the machine go, they should be able to use the blickets themselves to bring about effects on the machine. Second, children's causal reasoning involves both substantive prior knowledge, principles about what sorts of things cause what other things to happen, as well as formal principles about how patterns of conditional dependence and independence indicate causality. If children are using a causal map, they should combine both types of information to reach causal conclusions.

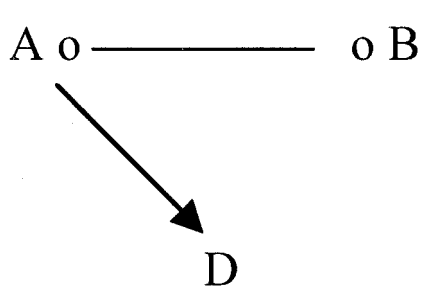


*Figure 11.* The causal structure of the blicket detector.

In the case of physical causality, and particularly in the case of machines, a likely general substantive principle is that if an event makes a machine go, the cessation of the event will make the machine stop—this applies to many common cases of switches, for example. This is not a necessary principle, of course, but it is a plausible pragmatic assumption about this type of causal relation. If children really think that the blickets have the causal power to make the machine go, they should also infer that removing the blicket is likely to make the machine stop, even if they have never seen this event. Moreover, they should intervene appropriately to bring about this effect. On the other hand, if they are merely associating the word, the object, and the effect, children should not draw this inference, nor should they be able to craft an appropriate intervention.

In a subsequent experiment (Gopnik et al., 2001, Experiment 3), we tested these ideas with 3- and 4-year-olds. We modified the task so that children did not see that removing the block made the machine stop (see Figure 12). The experimenter placed one block, B, on the machine, and nothing happened. The B block was removed, and then the experimenter placed the other block, A, on the machine, and the machine activated. After a few seconds (with the machine activating), she replaced the original B block on the machine next to the A block, and the machine continued to activate for an extended time. She then simply asked the children, "Can you make it stop?" If children were drawing causal conclusions from patterns of conditional dependence and independence and combining those conclusions with their substantive causal knowledge, they should remove the A block rather than the B block. We also used a similar two-cause control task. This involved exactly the same sequence of events, except that the B block did activate the machine when it was placed on the machine. In this case, children who use causal reasoning should remove both blocks.

The one-cause task always preceded the two-cause task, as this meant that children had never seen the right response at the time they made their intervention and so could not simply be imitating the experimenter. They had never seen that removing a block made the machine stop in the one-cause task or that removing both blocks made it stop in the two-cause task.

The children behaved as we predicted. In the one-cause task, they removed only Object A 75% of the time, significantly more often than any of the other responses (they removed Object B alone 12.5% of the time and removed both blocks simultaneously 12.5% of the time). Similarly, in the two-cause task, they removed both blocks simultaneously 50% of the time, significantly more often than they removed Object A alone (12.5% of the time) or Object B alone (27.5% of the time). Children were also significantly more likely to remove Object A in the one-cause task than in the two-cause task and were significantly more likely to remove both blocks simultaneously in the two-cause task than in the one-cause task.

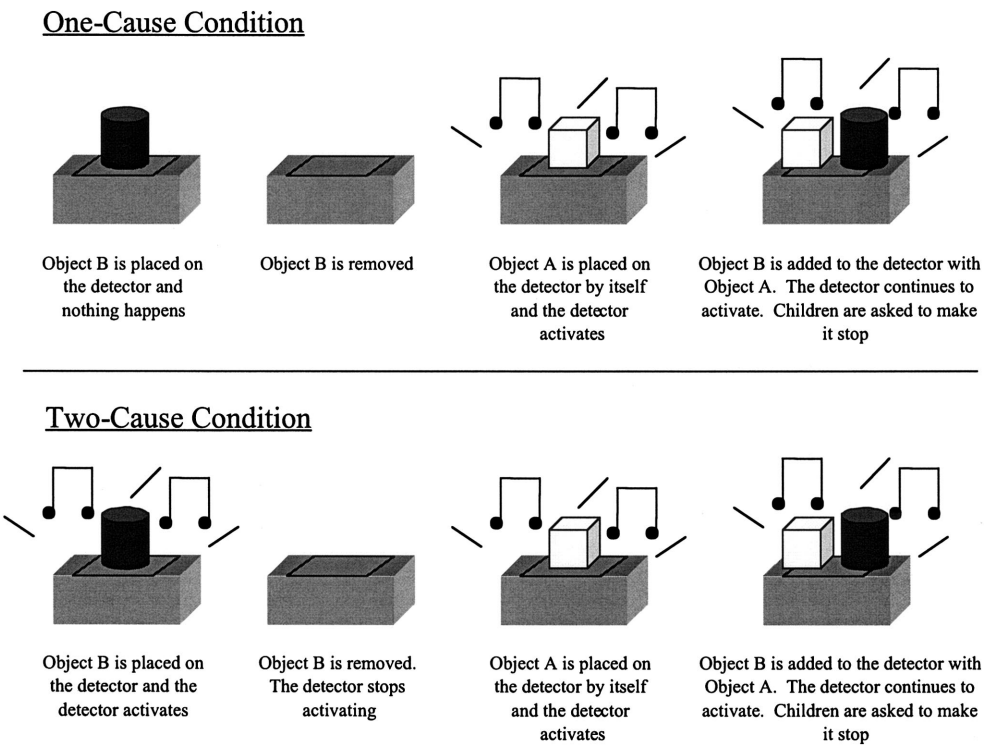


*Figure 12.* Procedure used in Gopnik et al. (2001, Experiment 3).

The results of these experiments rule out many possible hypotheses about children's causal learning. Because children did not activate the detector themselves, they could not have solved these tasks through operant conditioning or through trial-and-error learning. The blickets and nonblickets were perceptually indistinguishable, and both blocks were in contact with the detector, so children could not have solved the tasks through their substantive prior knowledge about everyday physics.

The "make it stop" condition in this experiment also showed that children's inferences went beyond classical conditioning, simple association, or simple imitative learning. Children not only associated the word and the effect, they combined their prior causal knowledge and the new causal knowledge they inferred from the dependencies to create a brand-new intervention that they had never witnessed before. As we mentioned above, this kind of novel intervention is the hallmark of a causal map. It is interesting that there is, to our knowledge, no equivalent of this result in the vast animal conditioning literature, although such an experiment would be easy to design. Would Pavlov's dogs, for example, intervene to silence a bell that led to shock, if they had simply experienced an association between the bell and the shock but had never intervened in this way before?

In all these respects, children seemed to have learned a new causal map. Moreover, this experiment showed that children were not using simple frequencies to determine the causal structure of this map but were using more complex patterns of conditional dependence. However, this experiment was consistent with all four learning models we described above, including the causal interpretation of the RW model.

*Inference from indirect evidence: Backward blocking.* In the next study we wanted to see whether children's reasoning would extend to even more complex types of conditional dependence and, in particular, if children would reason in ways that went beyond causal RW. There are a number of experimental results that argue against the RW model for adult human causal learning. One such phenomenon is "backward blocking" (Shanks, 1985; Shanks & Dickinson, 1987; Wasserman & Berglan, 1998). In backward blocking, learners decide whether an object causes an effect by using information from trials in which that object never appears.

Sobel and colleagues (Sobel, Tenenbaum, & Gopnik, in press) have demonstrated backward blocking empirically in young children. In one experiment (Sobel et al., in press, Experiment 2), 3- and 4-year-olds were introduced to the blicket detector in the same manner as in the Gopnik et al. (2001) experiments. They were told that some blocks were blickets and that blickets make the machine go. In a pretest, children saw that some blocks, but not others, made the machine go, and the active objects were labeled as blickets. Then children were shown two new blocks (A and B).

In one condition, the control, inference condition, A and B were placed on the detector together twice, and the detector responded both times. Then children observed that Object A did not activate the detector by itself. In the other condition, the backward blocking condition, children saw that two new blocks, A and B, activated the detector together twice. Then they observed that Block A did activate the detector by itself. In both conditions, children were then asked whether each block was a blicket and were asked to make the machine go (see Figure 13).

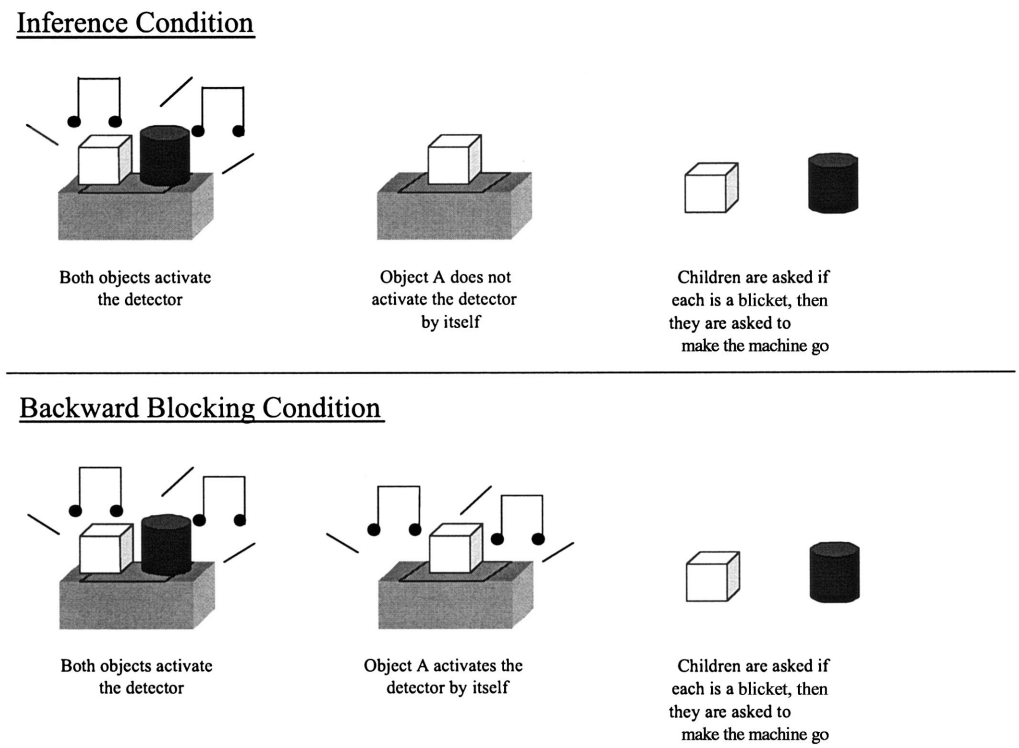


*Figure 13.* Procedure used in Sobel et al. (in press, Experiment 2).

In the control, inference condition, children said that Object A was a blicket only 6% of the time and always said that Object B was a blicket (100% of the time), significantly more often. Performance on the backward blocking condition was quite different: Children categorized Object A as a blicket 99% of the time. However, the critical question was how children would categorize Object B. Overall, children categorized Object B as a blicket only 31% of the time. In fact, even the youngest children categorized Object B as a blicket significantly less often in the backward blocking condition (50% of the time) than they did in the one-cause condition (100% of the time). In summary, children as young as 3 years old made different judgments about the causal power of Object B, depending on what happened with Object A. They used the information from trials that just involved A to make their judgment about B.

Children responded in a similar way to the "make it go" intervention question. This question was analogous to the "make it stop" question in Gopnik et al. (2001). Children had never seen the experimenter place the B block on the detector by itself in either condition. Nevertheless, in the inference condition they placed this block on the detector by itself 84% of the time. In the backward blocking condition they did so 19% of the time, significantly less often, and significantly less often than they placed the A block on the detector by itself (64% of the time).

What would the various learning models predict about this problem? In the pretest, children are shown that some blocks are blickets (about half the blocks, in fact). Children then have the following data in the following sequence.

Inference

1. A absent, B absent, E absent
2. A present, B present, E present
3. A present, B present, E present
4. A present, B absent, E absent

Backward blocking

1. A absent, B absent, E absent
2. A present, B present, E present
3. A present, B present, E present
4. A present, B absent, E present

According to the RW model, both *A* and *B* are positively associated with *E* (the effect). The last trial, Trial 4, should strengthen or weaken the association with *A* but should have no effect on the association with *B*, because *B* is absent. If that association is sufficiently strong, subjects should conclude that both *A* and *B* cause *E*. In particular, *B* should be equally strongly associated with *E* in the inference condition and the backward blocking condition.

In contrast, both Cheng's (1997) learning rule, with a suitable choice of focal sets, and constraint-based and Bayesian learning methods yield a qualitative difference between *A* and *B* in the backward blocking condition. In the RW model, the effect or lack of effect of the *A* block by itself has no influence on the judgment about *B*, but it has a crucial effect in these other models.

According to Cheng's (1997) methods, if the focal set for *A* in the backward blocking condition consists of Cases 1 and 4 (so *B*

has the same value in both cases—recall that a constant is independent, and conditionally independent, of everything), the estimate of the causal power of *A* is 1. If the focal set for *B* is Cases 2, 3, and 4 (so *A* has the same value, present, in all cases), the estimate of the causal power of *B* is undetermined.

By constraint-based reasoning, if the definition of conditional independence is $Pr(X, Y|Z) = Pr(X|Z) \times Pr(Y|Z)$ and the observed frequencies are representative of the probabilities (equivalently, the cases are multiplied by some large constant), *A* is not independent of *E* and is also not independent of *E* conditional on the absence of *B*. So, if there are no unobserved common causes, *A* is a cause of *E*. In contrast, *B* is associated with *E* but conditional on *A*, and also conditional on not *A*, *B* is independent of *E*. So, *B* is not a cause of *E*.

However, recall that constraint-based methods can also take uncertain prior knowledge into account, if inelegantly. In the current experiment, children know beforehand that about half the time blocks are blickets, and they seem to take this knowledge into account in their judgment about B.[3] That is, the Markov and faithfulness assumptions and the dependencies alone say that *A* is a cause but that *B* is not. Prior knowledge adds the fact that *B* may be a cause and is so about half the time.

Tenenbaum and Griffiths (2003) have recently described a Bayesian Bayes net learning model for backward blocking with the data above. This model more elegantly exploits the prior knowledge about blickets that was gained in the pretest. Their model rests on several assumptions: The presence of each block either is sufficient to cause *E* or is causally unrelated to *E*, the causal status of *A* and *B* are independent, and *E* does not occur unless *A* or *B* occurs. Bayesian updating then licenses the following inferences. After Cases 2 and 3, the posterior probability that *A* is a cause of *E* increases above its prior probability, and this is similarly true for *B*. But after Case 4, these quantities diverge. The probability that *A* is a cause of *E* becomes 1, because otherwise Case 4 could never be observed. The probability that *B* is a cause of *E* returns to its baseline prior probability, because the knowledge that *A* is surely a cause of *E* makes Cases 2 and 3, in retrospect, uninformative about the causal status of *B*.

All three of these models predict qualitative backward blocking. They predict a difference in the causal judgment about *B* in the two conditions, and RW does not. Constraint-based models with prior knowledge and Cheng (1997) methods predict that children will not judge that B is a blicket in the backward blocking case. The Bayesian model predicts, similarly but in more detail, that children will revert to their prior probability that B is a blicket, which is the guess they would make from the pretest evidence alone.

Children's categorization then was in qualitative accord with the Bayesian, constraint-based, and Cheng (1997) models but not with the RW model. It should, however, be noted that the RW learning rule can be modified to account for backward blocking by the addition of terms that decrease the association of a cue with an outcome when the outcome occurs in the absence of the cue (Wasserman & Berglan, 1998). It should also be noted that there is at least some evidence (Miller & Matute, 1996) that rats show something like backward blocking in the limited context of classical conditioning, which suggests that their inferences also go beyond RW. However, there is no evidence that animals can use backward blocking to create a new causal map and design a new intervention on the basis of that map, as the children in these experiments did.

### *Inferring the Direction of Causal Relations From Interventions and Correlations: The Puppet Machine*

The experiments we have described so far have shown that even very young children, as young as 30 months, are capable of some types of causal inference and learning that had previously been demonstrated only in human adults. We have also made the point that these types of inference and learning are consistent with the Bayes net representations and learning mechanisms. However, one of the principal advantages of applying a normative formalism such as Bayes nets is that it suggests new types of causal learning: types of causal learning that have not previously been explored in children or adults. The next three experiments use a new experimental paradigm to explore types of learning that are novel in three respects. First, they involve learning about the direction of the causal arrow between two simultaneous events: Did *X* cause *Y*, or did *Y* cause *X*? Second, they involve learning about this causal relation from a combination of observations of correlations and observations of interventions. Third, in the last experiment, they involve inferring an unobserved common cause of *X* and *Y*.

As we mentioned earlier, RW and other associationist causal learning procedures (including the modified rule that accounts for backward blocking) and the learning procedures in Cheng's (1997) published theory require information that distinguishes potential causes and effects. The potential causes and effects have to be discriminated before these models can operate. Then the models calculate individually either the association strength between each cause and each effect or the causal power of each cause to produce each effect. The models do not themselves generate the discrimination of causes and effects from observations of the dependencies. In contrast, Bayes net learning methods consider the fit between the entire data set and all the possible graphs that include the relevant variables, including graphs with arrows in one direction or the reverse direction.

One way Bayes net learning methods can solve the problem of simultaneous events is to use information about the effects of interventions. As we have repeatedly emphasized, one of the central advantages of the Bayes net formalism is that it gives a unified account of both correlations and interventions—it naturally explains how both predictions and interventions follow from causal structure, and it provides an account of how both observations of correlations and observations of interventions can be used in concert to infer causal structure. Work in the adult causal learning literature suggests that adults can learn about causal relations between events by looking at the patterns of correlation among those events. Though there are some suggestions in this adult literature about the treatment of interventions (Cheng, 1997,

[3] The prior knowledge explanation seems most plausible in this case. However, there is also another possible explanation for the uncertainty about *B*. As we noted in Footnote 2, it is possible that children used a difference formula for calculating the conditional probabilities (this is the method used in Cheng's, 1997, theory). According to this formula, the causal influence of *B* would not be calculable.

p. 375), there have not been systematic theoretical treatments or empirical investigations of this type of learning.[4]

On the other hand, work in the developmental literature, as well as work on operant conditioning and trial-and-error learning in animals, does suggest a limited kind of learning from intervention. As we mentioned earlier, this literature suggests that children (and animals) make the substantive assumption that their intentional actions cause the events that immediately follow those actions. The developmental literature on imitative learning also suggests that human children, at least, can make such inferences by observing the effects of the actions of others, though this is not so clear for animals. (Meltzoff & Prinz, 2002; Tomasello & Call, 1997).

The Bayes net formalism suggests ways children could combine information about interventions, their own or those of others, and information about events that are not the result of interventions to draw more complex causal conclusions. This includes conclusions about the causal relations between simultaneous events. Can children go beyond just assuming that their interventions directly cause the events that follow them? Can they also use information about the effects of interventions to discover the causal relationships among other events, as scientists do in experiments?

The blicket detector paradigm lies within the scope of the RW and Cheng (1997) models. The blickets are clearly the potential causes, and the light is the potential effect. Could a paradigm be designed that goes beyond the scope of these models? Schulz and colleagues (Schulz, 2001; Schulz & Gopnik, 2001) have designed such a paradigm: the puppet machine. We have used versions of this and similar techniques with both adults and kindergarteners (Kushnir, Gopnik, Schulz, & Danks, in press; Sobel & Kushnir, in press), but here we report only the results with 4-year-olds.

In these experiments, children saw two or three stylized "puppets," actually differently colored rubber balls attached to wooden sticks and placed behind a stage. The puppets were given different names on the basis of the color of the balls (e.g., "This is Reddy, and this is Bluey"). The puppets could be inserted into a single mechanism behind the stage, out of sight of the child, and the experimenter could move that mechanism up and down (invisibly) to move both puppets simultaneously, so that children simply saw correlations without interventions. The experimenter could also visibly intervene on each puppet separately by visibly pulling on the stick from above (see Figure 14).

*Experiment 1.* Sixteen 4-year-olds were tested. Children began with a pretest–training trial. Children saw the puppets move together and stop together simultaneously. They were told, "These two puppets move together and stop together. But one of these puppets is special. The special puppet always makes the other puppet move. Can you tell me which one is the special puppet?" Then the experimenter again (invisibly) made the puppets move together and stop together simultaneously. This time, while the puppets moved together, the experimenter explicitly identified the causal puppet, naming the puppets according to the color of their rubber balls. She said, "I'm moving X and X is making Y move. Which is the special puppet?" This gave the impression that one of the puppets had the power to make the other puppet move and that the experimenter was (invisibly) moving that special puppet, which then moved the other puppet (in fact, the experimenter was moving the common mechanism, but children did not know that). Children were only included in the experimental trials if they said that X was the special puppet. This training task established that special puppets caused other puppets to move.

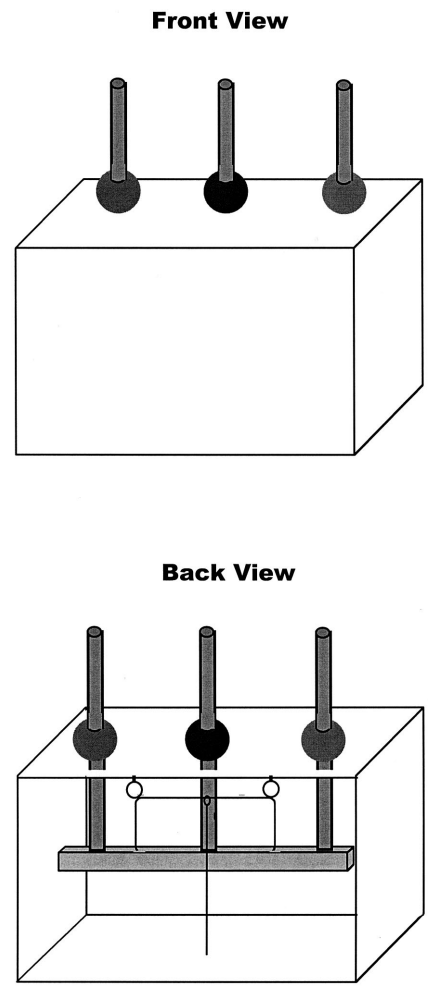


*Figure 14.* The puppet machine.

Then children received the experimental tasks, with new, differently colored puppets. In particular, we presented children with two types of tasks. In the *common effects* task the children first saw the puppets move together and stop together simultaneously four times. Then they saw the experimenter visibly intervene to make Y move by pulling on the stick from above, while X did not move. Then both puppets moved together again, and the experimenter asked, "Which is the special puppet?" The color and position of the special puppet were counterbalanced across trials.

[4] Very recently, Steyvers, Tenenbaum, Wagenmakers, and Blum (2003) have obtained results from an adult study, also explicitly inspired by Bayes net learning models, which in some respects parallels the studies we describe below with children. Their studies and ours were completed at the same time but independently. Adults were presented with dependencies among simultaneous events and had to infer causal direction. Either they simply observed the dependencies or they were allowed to experimentally intervene on the system. Adults were able to infer the direction of the relations to some extent just from observations, but their performance improved markedly when they were allowed to intervene.

The correct causal representation in this case is that the movement of Y is a common effect of the experimenters' intervention (which we will call *I*) and the movement of X: $I \rightarrow Y \leftarrow X$. If children infer this representation correctly they should conclude that X was special; it caused Y to move.

The second, *common cause* task involved three new differently colored puppets, X, Y, and Z. Children saw the puppets move together and stop together simultaneously several times. Then they saw the experimenter visibly intervene to make Y move by itself, while X and Z did not move, and saw the experimenter visibly intervene to make Z move by itself, while X and Y did not move. Again, they were asked to identify the special puppet. The correct causal representation in this case is that the movements of Y and Z are a common effect of the experimenter's interventions and X, and that X is a common cause of Y and Z. So, X is special; it caused Y and Z to move: $I_1 \rightarrow Z \leftarrow X \rightarrow Y \leftarrow I_2$.

In a control condition, 16 children of the same age saw a set of events that was similar perceptually, especially in terms of the salience of each puppet, but in which no causal conclusion was possible. The pretest, the test question, and the events of the puppets moving together and stopping together were the same, but the experimenter intervened by placing a rock in front of the Y, or Y and Z puppets instead of moving them.

Children received each of the tasks two times, with different puppets. Preschool children (as well as kindergarteners and undergraduates) made these inferences correctly. They chose X as the special puppet 78% of the time in the common effects tasks and 84% of the time in the common cause tasks, significantly more often than they chose the other puppet or puppets. In the control condition, children chose the X puppet 31% of the time and 34% of the time respectively, significantly less often than in the experimental condition, and no better than chance. Eleven out of the 16 children were correct on both common effects tasks, and 12 of the 16 were correct on both common cause tasks, significantly greater than chance in both cases.

Note especially that the children's causal judgments were about events—the motions of the puppets—for which they had no time order. Note also that they were not making inferences about the direct causal consequences of the experimenter's action. Instead, they used the action to make an inference about the causal relation between the two puppets.

Qualitatively, this result follows directly from the Markov and faithfulness assumptions. Applying the theory of intervention (described previously in this article) to the correct graphs generates the right predictions; applying it to the possible incorrect graphs does not (Pearl, 2000; Spirtes et al., 1993, 2001). There are three possible causal graphs of the observed variables given the instructions.

1. $X\ Y$.
2. $X \rightarrow Y$.
3. $Y \rightarrow X$.

The dependence between *X* and *Y* in the nonintervention trials eliminates Possibility 1, which implies that the two variables are independent. If Possibility 3 were the case then the intervention on *Y* would not cut the arrow between *Y* and *X*, as this arrow is not directed into the manipulated variable, and *X* and *Y* would still be dependent. In Possibility 2, however, the intervention does cut the arrow, so *X* and *Y* become independent in that case, always with the assumption that the data are representative of the probabilities. So, Possibility 2 is the only graph of the observed variables that is consistent with the instructions, the data, and the Markov and faithfulness assumptions. A similar reasoning process applies to the common cause case. The correct structures would be inferred by constraint-based and Bayesian learning methods. Because they require a prior specification of potential cause and effect, the causal RW rule and Cheng's (1997) power PC learning procedures do not make predictions in these cases.

*Experiment 2.* The next experiment was designed to eliminate an alternative explanation for Experiment 1 and to provide a direct contrast with the predictions of the causal RW account. In Experiment 1 children were told that the special puppet always made the other puppet move. Moreover, they had to make a forced choice between the two puppets. They were asked, "Which one is special?" which implied that one and only one puppet was special. In the common effects task it is possible that the children simply observed that Y moved and X did not on one trial and therefore rejected Y as a potential cause (because it did not always make the other puppet go). Because children were given a forced choice between the two alternatives, they said that X was the cause instead. They could have identified X as the special puppet merely because Y was no longer an option, not because they had inferred that causal structure from the pattern of dependencies.

However, if the experimenter did not specify that the causes were deterministic or ask for a forced choice, the fact that Y failed to cause X on one trial would not be grounds for eliminating Y as a candidate cause or saying that X was the cause instead. Y might usually but not always cause X, and neither or both of the puppets might be special. In Experiment 2, we modified the common effects procedure in this way. We told children that some puppets were special and that special puppets almost always make other puppets move. Instead of asking, "Which puppet is special?" we independently asked, "Is Y special? Does Y make X move?" and, "Is X special? Does X make Y move?" The Bayes net learning algorithms would still conclude that X causes Y and Y does not cause X (assuming, again, that the frequencies in the data are representative) because the pattern of dependencies is the same. They would generate the common effects causal structure, $I \rightarrow Y \leftarrow X$, as the correct causal graph given this pattern of evidence. Children should say "yes" to the question about X and "no" to the question about Y.

Eliminating the forced choice and allowing indeterministic causation also allowed us to directly contrast the predictions of Bayes net learning procedures and the causal RW learning procedure. As we discussed above, the causal RW procedure, including the procedure as modified to account for backward blocking, makes no prediction about how children might answer the question, "Which one is the cause?" because there are no procedures that distinguish causes from effects in this case. However, if children are asked, "Does X cause Y?" and, "Does Y cause X?" the question itself specifies the potential cause and potential effect. The causal RW model should predict that the answer to these questions will be based on the association strength between the potential cause and the potential effect. If there are a sufficient number of positive trials, then the answer should be "yes" to both questions.

This difference in predictions reflects the fact that causal RW considers the association strength between each potential cause and effect separately and does not distinguish between interventions and other events. In contrast, the Bayes net learning procedures assess the entire causal structure given the entire data pattern and treat interventions differently from other events.

However, contrasting these predictions depends on knowing whether, in fact, there is sufficient associative strength between Y and X to lead to a causal response using the RW rule. Note that X and Y are positively associated on all the trials in which X occurs. Y and X are positively associated on 5 out of the 6 trials in which Y occurs. Perhaps this difference is sufficient to lead children to say "yes" to X and "no" to Y. It could be that the single negative trial with Y weakens the association sufficiently to rule out a causal response.

In Experiment 2, we compared our modified version of the common effects puppet show with a new puppet show involving two new puppets, U and V. In the common effects puppet show, X and Y are perfectly associated except for a single trial on which Y moves but X does not. In the new association puppet show, two new puppets, U and V, are similarly perfectly associated except for a single trial on which U moves but V does not. However, in the new puppet show, the experimenter visibly intervenes on U in all the trials. On five trials, she pulls U up and down, and the other puppet, V, moves as well. On a single trial, however, the experimenter pulls U up and down, and V does not move. The question is whether this single negative trial will lead the children to rule out U as the cause of V.

Schematically, the two conditions can be represented as follows:

Common effects

1. Y and X
2. Y and X
3. Y and X
4. Y (by intervention) and −X
5. Y and X
6. Y and X

Association

1. U (by intervention) and V
2. U (by intervention) and V
3. U (by intervention) and V
4. U (by intervention) and −V
5. U (by intervention) and V
6. U (by intervention) and V

Note that the difference between the common effects condition and the association condition is entirely a difference in the balance of intervention and nonintervention trials. On the Bayes net account it is critical that one of the trials is an intervention and the others are not. On the associationist account the distinction between intervention trials and nonintervention trials is irrelevant—what matters for learning is the associations, however they are produced.

Thirty-two 4-year-old children (mean age = 4 years, 6 months) were randomly assigned to a common effects group of 16 children and an association group of 16 children. All the children in both groups began with a pretest–training test similar to that of Experiment 1 with two puppets, Z and Q. However, rather than telling the children, "One of these puppets is special. The special puppet always makes the other puppet move," the experimenter told the children, "Some of these puppets are special. Special puppets almost always make other puppets move." Moreover, instead of asking, "Which puppet is special?" the experimenter then asked, "Is Z special? Does Z make Q move?" To avoid setting a precedent for yes–no answers and thus implying a forced choice, the experimenter did not ask the children about Puppet Q in the pretest.

The procedure for the common effects group was just like the procedure in Experiment 1. Two new, differently colored puppets were placed in the first and third hole of the stage. The children saw the puppets move up and down together three times.

The experimenter then reached above the stage and grasped the dowel of Puppet Y and moved Y up and down within the child's sight. X did not move. Then the children again saw the puppets move up and down simultaneously twice in a row, with no visible intervention from the experimenter. The experimenter then asked the child, "Is X special? Does X make Y move?" and, "Is Y special? Does Y make X move?" The order of the questions and the location and color of the special puppet were counterbalanced between trials. If the children were able to understand the causal structure ($I \rightarrow Y \leftarrow X$) of the event, they should say that X is special and that Y is not special. The procedure was then repeated with two new puppets.

In the association condition, two new, differently colored puppets were placed in the first and third hole of the stage. This time, the experimenter manipulated one puppet visibly by moving the dowel from above and simultaneously (but surreptitiously) pulled the string behind the stage so that both puppets moved. From the child's point of view, the experimenter moved a puppet, which simultaneously made the second puppet move. The children saw the experimenter move Puppet U and saw Puppet V move too, three times in a row. On the next trial, however, the experimenter visibly moved U but did not surreptitiously pull the string, so V did not move. Then children saw the experimenter move Puppet U and saw Puppet V move simultaneously two more times. Finally, the experimenter asked, "Is U special? Does U make V move?" and, "Is V special? Does V make U move?" The order of the questions and the location and the color of the special puppet were counterbalanced between trials. The procedure was then repeated with two new puppets.

We first looked at children's overall performance. According to the Bayes net formalism, the correct answer to the two questions in the common effects task is that X is special and that Y is not special. In the association task, children should answer that U is special and that V is not special. Because they received two tasks in each condition, if children were performing at chance, they should show this pattern (that is, perform at ceiling) 6.25% of the time in each condition. In the common effects condition, 9 of the 16 children correctly identified X as the special puppet and Y as

not special across both trials, significantly more often than would be expected by chance ($p < .001$ by binomial test). Likewise, in the association condition, 11 of the 16 children correctly identified U as the special puppet and V as not special across both trials, significantly more than would be expected by chance ($p < .001$ by binomial test). The 9 out of 16 children performing at ceiling in the common effects condition in Experiment 2 is not significantly different from the 11 out of 16 children performing at ceiling in the common effects condition in Experiment 1, $\chi^2(1, N = 16) = 0.533$, *ns*, despite the fact that they had to answer two questions instead of one.

Thus, children's performance in the common effects task in Experiment 1 cannot be attributed to the forced-choice deterministic paradigm. Children were equally successful when they were asked to judge each puppet independently in the probabilistic context of Experiment 2. It might seem possible that the children in the common effects condition in Experiment 2 simply ignored the information in the training that special puppets probabilistically (almost always) make other puppets go and still ruled out the Y puppet because of the single negative trial. However, the results of the association condition rule out this possibility. Children chose U as a special puppet significantly above chance despite the fact that U also failed to make the other puppet move on one trial. The results of the association condition also suggest that children do not require causes to behave deterministically; they said U caused V to move, even though it produced the expected effect only five sixths of the time.

We can also consider the implications of these results for the causal RW account. The crucial comparison concerns the children's responses to the Y and U puppets. According to the Bayes net learning procedures, children should say that Puppet Y is not special and that Puppet U is special; according to the RW procedures, children should say that both Puppet Y and Puppet U are special (or, less probably, that they are both not special). If children were performing at chance, children should show the expected pattern across the two trials 25% of the time.

On the two trials of the common effects condition, 10 children consistently said that Y was not special, significantly more than the 1 child in the association condition who consistently said that U was not special, $\chi^2(1, N = 16) = 11.22$, $p < .001$. Similarly, 11 children in the association condition consistently said that U was special, significantly more than the 5 children in the common effects condition who consistently said that Y was special, $\chi^2(1, N = 16) = 4.50$, $p < .05$. As predicted by the Bayes net learning procedures but not by the RW procedures, the pattern of responses to the Y puppet in the common effects condition differs significantly from the pattern of responses to the U puppet in the association condition, $\chi^2(1, N = 16) = 8.43$, $p < .01$.

These results suggest that the causal RW learning mechanism cannot fully account for children's inferences about causal structure. The association between Y and X and between U and V was identical in the common effects and the association conditions. Yet when Y occurred without X in the common effects case, children correctly inferred that Y did not cause X. When U occurred without V in the association case, children correctly inferred that U did cause V.

*Experiment 3: Unobserved causes.* As we mentioned earlier, the causal RW learning rules and the learning rule in Cheng (1997) can only be applied correctly when there are no unobserved common causes of the cause and the effect (although a generalization of Cheng's rule can be applied in some such cases; see Glymour, 2001). In particular, associationist causal learning models, such as causal RW, do not permit learners to distinguish between cases in which associations occur because events of one type cause events of the other type and cases in which associations occur because events of each type are influenced by events of a third kind, which is unobserved. The puppet machine paradigm provides us with a way to test whether children will attribute causal influence to a common unobserved variable.

Notice that the graphical structure in the three-puppet common cause task we described above is the same as the structure in Figure 9 in our discussion of unobserved variables. A simple modification of this task, involving just two puppets, allows us to test whether children will infer unobserved common causes. We have completed a first preliminary study using this task.

In this task, sixteen 4.5-year-old children (mean age = 4 years, 10 months) received the same pretest and training trials as those in Experiment 1, with deterministic causal relations (i.e., "The special puppet always makes the other puppet go"). However, rather than just asking which puppet was special or asking whether each individual puppet was special, the experimenter asked the children to explain the events: She asked, "Why are the puppets moving together?" Then children received a common effects task. This task proceeded in exactly the same way as the task in Experiment 1, except that children were asked to explain why the puppets were moving together rather than being asked which puppet was special or whether each puppet individually was special. (If children refused to answer the explanation question spontaneously, we presented them with a choice: "Is it X, is it Y, or is it something else?")

The children were then presented with an unobserved variable task with two new puppets. The children first saw both puppets move together on four trials, as in the previous tasks. Then the experimenter visibly intervened to move Y, and X remained unmoved, as in the common effect experiment. But this time the experimenter then also visibly intervened to move X, and Y remained unmoved. Finally, children again saw both puppets move together. Again, children were simply asked, "Why are the puppets moving together?" (with the additional choice of three options for the children who refused to answer).

Given the instructions, the Markov and faithfulness assumptions, and the data, this pattern of events is incompatible with any of the three acyclic graphs involving just $X$ and $Y$: $X\ Y$, $X \rightarrow Y$, or $Y \rightarrow X$ (see discussion, above). This means that the only way to explain the data is to postulate a graph with an unobserved common cause of $X$ and $Y$, $X \leftarrow U \rightarrow Y$. Children should conclude that some other unobserved common cause (most obviously, the experimenter behind the screen) is responsible for the trials in which the objects move together.

In the common effects trials, all 16 children chose one puppet as the explanation for the movement (e.g., "X is pushing Y," "X is making Y move"). Thirteen of the 16 children chose the correct puppet, similar to their performance in the earlier experiments. Thus, a majority of 4-year-olds solved the common effects task across all three experiments, whether they were asked to choose between the puppets, to identify whether each puppet was special, or to explain the puppets' movement.

However, a majority of children in the unobserved condition posited an unobserved common cause. Nine of the 16 children said that some unobserved variable (e.g., "your hands," "you behind there," "something else") was responsible for the puppets moving together, significantly greater than the zero children who said this in the common effects condition. Only 5 of the 16 children said that one puppet had caused the movement, significantly fewer than the 16 children who said this in the common effects condition. (The remaining 2 children refused to answer even after they were presented with the three options.) These children postulated that observed variables were causes, unless no graph was consistent with the dependencies between those variables. In that case, they postulated an unobserved common cause.

### *Further Experiments*

These experiments are only a first step, and our results need to be further replicated with additional controls. Two control conditions would be particularly compelling. One would be to do the tasks in Experiments 2 and 3 but to have the experimenter point to each object as it moves rather than intervening to make the objects move. This task would be almost identical to the original tasks perceptually, in terms of salience, and in terms of any measure of associative strength, but we predict that children would not judge causal influence in the same way, because of the lack of interventions. A second interesting experiment would be to show children a pattern of independence rather than dependence in the nonintervention trials—that is, to show them puppets moving at random with respect to one another, followed by either one or two failed interventions. We predict that in this case children would conclude that the puppets are moving because of two independent unobserved causes rather than because one is moving the other or because both are moved by an unobserved common cause. We have performed these experiments with adults, and our predictions were confirmed (Kushnir et al., in press). Work with children is ongoing.

These experiments also do not definitively tell us which causal learning methods children use. They do suggest that these methods extend beyond the causal RW rule or Cheng's (1997) learning rule as published and that they yield results that are normatively correct according to the Bayes net formalism. As we said earlier, however, we do not believe that children's actual learning methods are as general as the Bayesian or constraint-based algorithms, and we need to discover empirically where the limits lie. Perhaps current versions of associationist models or Cheng models could be modified, supplemented, and extended to produce comparable results.[5] However, such a project would prove the value of the Bayes net formalism and justify the existence of this article. The formalism suggests causal learning problems, such as the simultaneous events problem or the unobserved common cause problem, that have not been explored before, and it provides normatively accurate solutions to these problems. Children's behavior can be assessed in the light of these solutions, and more specialized learning proposals can be judged and modified in terms of this more general framework.

Moreover, dozens of other such experimental questions suggest themselves. The Bayes net formalism allows predictions about the inference of more complex causal structures, such as inhibitory causes or interactive causes, and it allows us to make inferences about causal chains, where *X* causes *Y*, which causes *Z*. We could explore whether children will make accurate inferences in these cases. Bayes nets are designed to deal with probabilistic as well as deterministic data, and Bayes net applications almost always involve probabilistic data. Two of our experiments suggest that children make causal inferences even in nondeterministic cases. In the two-cause condition of Gopnik et al. (2001), children thought the object that set the machine off two out of three times was a blicket, and in the association condition of the puppet experiment, they said that the puppet that activates the other puppet five of six times was special. We could explore how children will reason about probabilistic causal relations as well as deterministic relations. The formalism provides ways of combining information about conditional probabilities with substantive prior knowledge, such as knowledge of time order. The "make it stop" experiment showed that children can combine formal assumptions about causality with substantive prior causal knowledge. We can see how children combine specific kinds of prior knowledge with this kind of reasoning.

### *Further Computational Work*

Just as the Bayes net formalism suggests new experiments with children, applying Bayes nets to children's cognition suggests new computational questions. The formalism has mostly been applied in practice to "data-mining" problems. These problems are unlike the problems that children face in many respects. As it currently stands, the learning models that use the formalism do not specify how we decide that particular events are instances of a variable, nor how we decide which variables to consider. Nor do they propose exact learning processes that have psychologically plausible memory and processing limitations.

It seems extremely unlikely, for example, that children store vast quantities of data in memory and then apply learning procedures to the data. Instead, they must surely form hypotheses and use them on the basis of small samples of data, forget the data or

[5] For example, Patricia Cheng (personal communication, December 2002) has suggested the following account of the common effect puppet experiments: The instructions imply that either A → B or B → A. Ruling out the former on the basis of the intervention therefore yields the latter by deduction. This conclusion is consistent with the retroactive application of the causal power equation at this point (i.e., after the intervention trial and the ruling out of the A → B causal direction) to the evaluation of B → A in the nonintervention trials. The intervention trial is excluded in this focal set because with respect to B, the candidate in question, I, is an alternative cause and needs to be kept constant, constantly absent in this case because that is the only way to satisfy the independent-occurrence assumption. If a subject accepts the assumption implied in the instructions that no unobserved cause is assumed to exist (so that either A or B is special), then in the evaluation of B → A, alternative causes occur independently of B (trivially, because of the assumption that there are no other causes), and the causal power equation gives $qBA = (1 - 0)/(1 - 0) = 1$ (where qBA is the probability that B causes A given that B occurs, as in Cheng, 1997). But, if subjects refuse to accept that assumption and instead allow for the possibility of an unobserved cause, then they would be uncertain whether $B \rightarrow A$: Alternative causes may or may not occur independently of *B* in the nonintervention trials. Both answers, Cheng argues, can be explained by her theory, depending on how a subject interprets the experimenters' instructions.

most of them, and revise their hypotheses as required by new data. In the course of this revision they quite possibly alter not only the causal relations they hypothesize but also the variables and properties they consider to be useful. Equally, causal regularities that are learned for one context may somehow constrain the causal regularities to be learned in other contexts, especially when the domains overlap in objects or properties, allowing for a kind of learning by analogy.

Moreover, it seems possible, and even likely, that children begin the process of causal learning with some innately given assumptions about which variables are relevant to causal inference, and perhaps with some assumptions about structure, such as assumptions that some variables are causally connected to others. Children might be born assuming at least some sketchy causal graphs. These would correspond to the innate "starting-state" theories proposed by some "theory theorists" (see Gopnik & Meltzoff, 1997; Gopnik & Wellman, 1994). However, these initial assumptions could be modified or overturned in the light of new data about conditional dependencies and interventions. Children, then, may be bootstrappers as much as data miners.

There is computational research under way on all of these issues, but it is as yet far from providing a firm understanding of the possibilities. Glymour (2001) suggested a number of heuristics for transforming a given set of variables in a network to new ones; Spirtes (2001) has shown that certain constraint-based algorithms have the property that they give incomplete but correct information if they are stopped (e.g., because of limits on computational resources or time) at any point in the procedure. Bayesian learning algorithms can store the best Bayes net found to explain a data set, and forgetting that data set, use that best hypothesis as a starting point for a greedy search if the hypothesis is to be revised in the light of new data.

## Conclusion

We end by returning to the analogy with vision science. At its beginnings in the 17th century, science emerged from two separate enterprises: natural history, which catalogued and described the world, and natural philosophy, which sought to explain the world in mathematical terms. For 100 years, vision science was primarily concerned with psychological natural history, discovering consistent patterns in visual experience and relating those patterns to the structure of objects in the world. These psychological findings were an absolutely necessary precursor to the current computational and neurological work. Without psychophysics and perceptual psychology, there would be no vision science. More recently, however, psychologists have gained a new and deeper understanding of vision by relating these psychological findings to a natural philosophy that tells them, in computational terms, how it is possible for the perceptual system to recover accurate information about the world.

The last 30 years have been a golden age for the natural history of children's learning. We know more than ever before about the consistent patterns in children's conceptions of the world and the consistent changes in those conceptions. But there has been much less natural philosophy; we have known much less about how it is possible for children to learn as much as they do about the world around them. Our hope is that just as perceptual psychology led to vision science, cognitive developmental psychology will be the first step toward a new learning science.

## References


Ahn, W., Gelman, S. A., Amsterlaw, J. A., Hohenstein, J., & Kalish, C. W. (2000). Causal status effects in children's categorization. *Cognition, 76,* 35–43.

Atran, S. (1990). *Cognitive foundations of natural history: Towards an anthropology of science.* New York: Cambridge University Press.

Bartsch, K., & Wellman, H. M. (1995). *Children talk about the mind.* New York: Oxford University Press.

Bullock, M., Gelman, R., & 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.

Campbell, J. (1995). *Past, space and self.* Cambridge, MA: MIT Press.

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

Cartwright, N. (1989). *Nature's capacities and their measurement.* New York: Clarendon Press.

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

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

Danks, D. (2003). Equilibria of the Rescorla–Wagner model. *Journal of Mathematical Psychology, 46,* 109–121.

Danks, D., Tenenbaum, J., & Griffiths, T. (2003). Dynamical causal learning. In S. Becker, S. Thrun, & K. Obermayer (Eds.), *Advances in neural information processing systems* (Vol. 15, pp. 67–74). Cambridge, MA: MIT Press.

Elman, J. L., Bates, E. A., Johnson, M. H., & Karmiloff-Smith, A. (1996). *Rethinking innateness: A connectionist perspective on development.* Cambridge, MA: MIT Press.

Flavell, J. H., Green, F. L., & Flavell, E. R. (1995). Young children's knowledge about thinking. *Monographs of the Society for Research in Child Development, 60,* v–96.

Gallistel, C. R. (1990). *The organization of learning.* Cambridge, MA: MIT Press.

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

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

Glymour, C., & Cheng, P. (1999). Causal mechanism and probability: A normative approach. In K. Oaksford & N. Chater (Eds.), *Rational models of cognition* (pp. 295–313). Oxford, England: Oxford University Press.

Glymour, C., & Cooper, G. (1999). *Computation, causation, and discovery.* Menlo Park, CA: AAAI/MIT Press.

Gopnik, A. (1988). Conceptual and semantic development as theory change. *Mind and Language, 3,* 163–179.

Gopnik, A. (2000). Explanation as orgasm and the drive for causal understanding: The evolution, function and phenomenology of the theory-formation system. In F. Keil & R. Wilson (Eds.), *Cognition and explanation* (pp. 299–323). Cambridge, MA: MIT Press.

Gopnik, A., & Esterly, J. (1999, April). *Causal inferences about material kinds.* Poster presented at the Meeting of the Society for Research in Child Development, Albuquerque, NM.

Gopnik, A., & Glymour C. (2002). Causal maps and Bayes nets: A cognitive and computational account of theory-formation. In P. Carruthers, S. Stich, & M. Siegal (Eds.), *The cognitive basis of science* (pp. 117–132). Cambridge, England: Cambridge University Press.

Gopnik, A., & Meltzoff, A. (1997). *Words, thoughts and theories.* Cambridge, MA: MIT Press.

Gopnik, A., & Nazzi, T. (2003). Words, kinds and causal powers: A theory

perspective on early naming and categorization. In D. Rakison & L. Oakes (Eds.), *Early categorization* (pp. 303–329). Oxford, England: Oxford University Press.

Gopnik, A., & Sobel, D. M. (2000). Detecting blickets: How young children use information about causal properties in categorization and induction. *Child Development, 71,* 1205–1222.

Gopnik, A., Sobel, D. M., Schulz, L., & Glymour, C. (2001). Causal learning mechanisms in very young children: Two, three, and four-year-olds infer causal relations from patterns of variation and covariation. *Developmental Psychology, 37,* 620–629.

Gopnik, A., & Wellman, H. M. (1994). The theory theory. In L. Hirschfield & S. Gelman (Eds.), *Mapping the mind: Domain specificity in cognition and culture* (pp. 257–293). New York: Cambridge University Press.

Harris, P. L., German, T., & Mills, P. (1996). Children's use of counterfactual thinking in causal reasoning. *Cognition, 61,* 233–259.

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

Heckerman, D. (1995). A Bayesian approach to learning causal networks (Tech. Rep. MSR-TR-95–04). San Francisco: Microsoft Research.

Heckerman, D., Meek, C., & Cooper, G. (1999). A Bayesian approach to causal discovery. In C. Glymour & G. Cooper (Eds.), *Computation, causation, and discovery* (pp. 143–167). Cambridge, MA: MIT Press.

Heider, F. (1958). *The psychology of interpersonal relations.* New York: Wiley.

Hickling, A. K., & Wellman, H. M. (2001). The emergence of children's causal explanations and theories: Evidence from everyday conversation. *Developmental Psychology, 5,* 668–683.

Hume, D. (1978). *A treatise of human nature.* Oxford, England: Oxford University Press. (Original work published 1739)

Inagaki, K., & Hatano, G. (1993). Young children's understanding of the mind body distinction. *Child Development, 64,* 1534–1549.

Jordan, M. (Ed.). (1998). *Learning in graphical models.* Cambridge, MA: MIT Press.

Kalish, C. W. (1996). Preschoolers' understanding of germs as invisible mechanisms. *Cognitive Development, 11,* 83–106.

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

Kiiveri, H., & Speed, T. (1982). Structural analysis of multivariate data: A review. In S. Leinhardt (Ed.), *Sociological methodology* (pp. 209–220). San Francisco: Jossey-Bass.

Kushnir, T., Gopnik, A., Schulz, L., & Danks, D. (in press). Inferring hidden causes. In *Proceedings of the twenty-fifth annual conference of the Cognitive Science Society.* Mahwah, NJ: Erlbaum.

Lagnado, D., & Sloman, S. A. (2002). Learning causal structure. In W. Gray & C. Schunn (Eds.), *Proceedings of the twenty-fourth annual conference of the Cognitive Science Society* (pp. 560–565). Mahwah, NJ: Erlbaum.

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

Leslie, A., & Roth, D. (1993). What autism teaches us about metarepresentation. In S. Baron-Cohen, H. Tager-Flusberg, & D. J. Cohen (Eds.), *Understanding other minds: Perspectives from autism* (pp. 83–111). New York: Oxford University Press.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information.* San Francisco: Freeman.

Mayhew, J. E., & Longuet-Higgins, H. C. (1982, June 3). A computational model of binocular depth perception. *Nature, 5865,* 376–378.

Meek, C. (1995). Strong completeness and faithfulness in Bayesian networks. In P. Besnard & S. Hanks (Eds.), *Uncertainty in artificial intelligence: Proceedings of the eleventh conference* (pp. 411–418). San Francisco: Morgan Kaufmann.

Meltzoff, A. N. (1988a). Infant imitation after a 1-week delay: Long-term memory for novel acts and multiple stimuli. *Developmental Psychology, 24,* 470–476.

Meltzoff, A. N. (1988b). Infant imitation and memory: Nine-month-olds in immediate and deferred tests. *Child Development, 59,* 217–225.

Meltzoff, A. N., & Prinz, W. (Eds.). (2002). *The imitative mind: Development, evolution, and brain bases.* Cambridge, England: Cambridge University Press.

Michotte, A. E. (1962). *Causalite, permanence et realite phenomenales; etudes de psychologie experimentale* [*Phenomenal causality, permanence and reality*]. Oxford, England: Publications Universitaires.

Miller, R. R., & Matute, H. (1996). Biological significance in forward and backward blocking: Resolution of a discrepancy between animal conditioning and human causal judgment. *Journal of Experimental Psychology: General, 125,* 370–386.

Nazzi, T., & Gopnik, A. (2000). A shift in children's use of perceptual and causal cues to categorization. *Developmental Science, 3,* 389–396.

Novick, L., & Cheng, P. (in press). Assessing interactive causal influence. *Psychological Review.*

Oakes, L. M., & Cohen, L. B. (1990). Infant perception of a causal event. *Cognitive Development, 5,* 193–207.

O'Keefe, J., & Nadel, L. (1978). *The hippocampus as a cognitive map.* New York: Oxford University Press.

Palmer, S. (1999). *Vision science: From photons to phenomenology.* Cambridge, MA: MIT Press.

Palmerino, C. C., Rusiniak, K. W., & Garcia, J. (1980, May 16). Flavor-illness aversions: The peculiar roles of odor and taste in memory for poison. *Science, 208,* 753–755.

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

Pearl, J. (2000). *Causality.* New York: Oxford University Press.

Perner, J. (1991). *Understanding the representational mind.* Cambridge, MA: MIT Press.

Piaget, J. (1929). *The child's conception of the world.* New York: Harcourt, Brace.

Piaget, J. (1930). *The child's conception of physical causality.* New York: Harcourt, Brace.

Povinelli, D. J. (2001). *Folk physics for apes.* New York: Oxford University Press.

Ramsey, J., Roush, T., Gazis, P., & Glymour, C. (2002). Automated remote sensing with near-infra-red reflectance spectra: Carbonate recognition. *Data Mining and Knowledge Discovery, 6,* 277–293.

Rehder, B., & Hastie, R. (2001). Causal knowledge and categories: The effects of causal beliefs on categorization, induction, and similarity. *Journal of Experimental Psychology: General, 3,* 323–360.

Reichenbach, H. (1956). *The direction of time.* Berkeley, CA: University of California Press.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical conditioning II: Current theory and research* (pp. 64–99). New York: Appleton-Century-Crofts.

Richardson, T. (1996). Discovering cyclic causal structure (Tech. Rep. CMU Phil 68). Pittsburgh, PA: Carnegie Mellon University.

Rogers, B. J., & Bradshaw, M. F. (1993, January 21). Vertical disparities, differential perspective and binocular stereopsis. *Nature, 6409,* 253–255.

Rovee-Collier, C. (1987). Learning and memory in infancy. In J. D. Osofsky (Ed.), *Handbook of infant development* (2nd ed., pp. 139–168). Oxford, England: Wiley.

Salmon, W. (1984). *Scientific explanation and the causal structure of the world.* Princeton, NJ: Princeton University Press.

Scheines, R., Spirtes, P., Glymour, C., & Meek, C. (1994). *TETRAD II.* Hillsdale, NJ: Erlbaum.

Schulz, L. (2001, October). *Do-calculus: Inferring causal relations from*

*observations and interventions.* Paper presented at the meeting of the Cognitive Development Society, Virginia Beach, VA.

Schulz, L., & Gopnik, A. (2001, December). *Inferring causal relations from observations and interventions.* Paper presented at the meeting of the Causal Inference Workshop: The Neural Information Processing, Whistler, British Columbia, Canada.

Shanks, D. R. (1985). Forward and backward blocking in human contingency judgment. *Quarterly Journal of Experimental Psychology: Comparative and Physiological Psychology, 37*(B), 1–21.

Shanks, D. R., & Dickinson, A. (1987). Associative accounts of causality judgment. In G. H. Bower (Ed.), *The psychology of learning and motivation: Advances in research and theory* (Vol. 21, pp. 229–261). San Diego, CA: Academic Press.

Shipley, B. (2000). *Cause and correlation in biology.* Oxford, England: Oxford University Press.

Shultz, T. R. (1982). Rules of causal attribution. *Monographs of the Society for Research in Child Development, 47*(Serial No. 194).

Slaughter, V., & Gopnik, A. (1996). Conceptual coherence in the child's theory of mind: Training children to understand belief. *Child Development, 67,* 2967–2988.

Slaughter, V., Jaakkola, R., & Carey, S. (1999). Constructing a coherent theory: Children's biological understanding of life and death. In M. Siegal & C. Peterson (Eds.), *Children's understanding of biology and health* (pp. 71–96). Cambridge MA: Cambridge University Press.

Sloman, S. A., & Lagnado, D. (2002). Counterfactual undoing in deterministic causal reasoning. In W. Gray & C. Schunn (Eds.), *Proceedings of the twenty-fourth annual conference of the Cognitive Science Society* (pp. 828–833). Mahwah, NJ: Erlbaum.

Sobel, D. M., & Gopnik, A. (2003). *Causal prediction and counterfactual reasoning in young children: Separate or similar processes?* Manuscript submitted for publication, Brown University.

Sobel, D. M., & Kushnir, T. (in press). Interventions do not solely benefit causal learning: Being told what to do results in worse learning than doing it yourself. In *Proceedings of the twenty-fifth annual meeting of the Cognitive Science Society.* Mahwah, NJ: Erlbaum.

Sobel, D. M., Tenenbaum, J. B., & Gopnik, A. (in press). Children's causal inferences from indirect evidence: Backwards blocking and Bayesian reasoning in preschoolers. *Cognitive Science.*

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

Spellman, B. A. (1996). Acting as intuitive scientists: Contingency judgments are made while controlling for alternative potential causes. *Psychological Science, 7,* 337–342.

Spirtes, P. (2001). An anytime algorithm for causal inference. In T. Jaakkola & T. Richardson (Eds.), *Proceedings of AISTATS* (pp. 213–231). San Francisco: Morgan Kaufmann.

Spirtes, P., Glymour, C., & Scheines, R. (1993). *Causation, prediction, and search* (Springer lecture notes in statistics). New York: Springer-Verlag.

Spirtes, P., Glymour, C., & Scheines, R. (2001). *Causation, prediction, and search* (Springer lecture notes in statistics, 2nd ed., rev.). Cambridge, MA: MIT Press.

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

Tenenbaum, J. B., & Griffiths, T. L. (2003). Theory-based causal inference. In S. Becker, S. Thrun, & K. Obermayer (Eds.), *Proceedings of the 2002 Neural Information Processing Systems Conference* (Vol. 15, pp. 35–42). Cambridge, MA: MIT Press.

Thelen, E., & Smith, L. B. (1994). *A dynamic systems approach to the development of cognition and action.* Cambridge, MA: MIT Press.

Tolman, E. C. (1932). *Purposive behavior in animals and men.* New York: Century.

Tomasello, M., & Call, J. (1997). *Primate cognition.* New York: Oxford University Press.

Waldmann, M. R., & Hagmayer, Y. (2001). Estimating causal strength: The role of structural knowledge and processing effort. *Cognition, 1,* 27–58.

Waldmann, M. R., & Martignon, L. (1998). A Bayesian network model of causal learning. In M. A. Gernsbacher & S. J. Derry (Eds.), *Proceedings of the twentieth annual conference of the Cognitive Science Society* (pp. 1102–1107). Mahwah, NJ: Erlbaum.

Wasserman, E. A., & Berglan, L. R (1998). Backward blocking and recovery from overshadowing in human causal judgment: The role of within-compound associations. *Quarterly Journal of Experimental Psychology: Comparative and Physiological Psychology, 51,* 121–138.

Watson, J. S., & Ramey, C. T. (1987). Reactions to response-contingent stimulation in early infancy. In J. Oates & S. Sheldon (Eds.), *Cognitive development in infancy* (pp. 77–85). Hillsdale, NJ: Erlbaum.

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

Wellman, H. M., Hickling, A. K., & Schult, C. A. (1997). Young children's psychological, physical, and biological explanations. In H. M. Wellman & K. Inagaki (Eds.), *The emergence of core domains of thought: Children's reasoning about physical, psychological, and biological phenomena* (pp. 7–25). San Francisco, CA: Jossey-Bass.

Received May 14, 2001
Revision received September 24, 2002
Accepted October 28, 2002 ■

**E-Mail Notification of Your Latest Issue Online!**

Would you like to know when the next issue of your favorite APA journal will be available online? This service is now available to you. Sign up at http://watson.apa.org/notify/ and you will be notified by e-mail when issues of interest to you become available!