In Gopnik, Alison; Schulz, Laura (Eds.) (2007). *Causal learning: Psychology, philosophy, and computation*. (pp. 67-85). New York: Oxford University Press. 358 pp.

# Learning From Doing: Intervention and Causal Inference

Laura Schulz, Tamar Kushnir & Alison Gopnik

*There is something fascinating about science. One gets such wholesale returns of conjecture out of such a trifling investment of fact.* ~ Mark Twain, 1883

Twain meant it as a witticism of course but there *is* something fascinating about science. From a few bones, scientists infer the existence of dinosaurs, from a few spectral lines, the composition of nebulae, and from a few fruit flies, the mechanisms of heredity. From a similarly trifling investment, some of us presume to conjecture even about the mechanisms of conjecture itself.

Why does science, at least some of the time, succeed? Why does it generate accurate predictions and effective interventions? With due respect for our accomplished colleagues, we believe it may be because getting wholesale returns out of minimal data is a commonplace feature of human cognition. Indeed, we believe the most fascinating thing about science may be its connection to human learning in general, and in particular, to the rapid, dramatic, learning that takes place in early childhood. This view, the *theory theory*, suggests that starting in infancy, continuing through the life span, and canalized in scientific inquiry, many aspects of human learning can be best explained in terms of theory formation and theory change.

Theories have been described with respect to their structural, functional, and dynamic properties (Gopnik & Meltzoff, 1997). Thanks to several decades of work in developmental psychology, we now know a great deal about the structural and functional aspects of children's theories. That is, in many domains, we know that children have abstract, coherent, causal, representations of events, we know something about the content of those representations, and we know what types of inferences they support.

We know for instance, that 6-month-olds' naïve physics includes principles of cohesion, continuity and contact, but not the details of support relations (Baillargeon, Kotovsky, & Needham, 1995; Spelke, Breinlinger, Macomber, & Jacobson, 1992; Spelke, et al., 1994). We know that 4-year-olds' naïve biology supports inferences about growth, inheritance and illness but not the adult concept of 'living thing' or 'alive' (Carey, 1985; Gelman & Wellman, 1991; Inagaki & Hatano, 1993; Kalish, 1996). We know that 2-year-olds' naïve psychology includes the concepts of intention and desire but not the concept of belief (Flavell, Green, & Flavell, 1995; Gopnik & Wellman, 1994; Perner, 1991). Moreover, we know that across domains, children's naïve theories support coherent predictions, explanations and even counterfactual claims (Harris, German, & Mills, 1996; Sobel, 2001; Wellman, Hickling, & Schult, 1997).

However, the theory theory is not just a theory about what children know or what children can do. It is, centrally, a claim about how children learn. In this respect, it is the dynamic rather than the structural and functional aspect of theories that is critical. If children's reasoning is like scientific theory formation, then children's naïve theories should be subject to confirmation, revision and refutation and children should be able to make inferences based on evidence from observation, experimentation and combinations

of the two.

Until recently, this dynamic feature of theories has been difficult to explain. If children's knowledge about the world takes the form of naïve theories -- and if conceptual development in childhood is analogous to theory change in science -- then we would expect the causal reasoning of even very young children to be very sophisticated. A causal "theory" (as distinct from, for instance, a causal module or a causal script) must support novel predictions and interventions, account for a wide range of data, enable inferences about the existence of unobserved and even unobservable causes, and change flexibly with evidence (Gopnik & Meltzoff, 1997). Moreover, theories have a complex relationship with evidence; they must be defeasible in the face of counter-evidence – but they can't be too defeasible. Because evidence is sometimes misleading and sometimes fails to be representative, the process of theory formation must be at once conservative and flexible.

In recent work we have focused on causal learning as a fundamental dynamic mechanism underlying theory formation. In thinking about what causal knowledge is we have been influenced by recent philosophical and computational work proposing an "interventionist" view of causation (see Woodward, Hitchcock, and Campbell this volume). This view stands in contrast to many traditional ideas about causation in both adult and developmental psychology. However, we believe that an interventionist account of causation not only helps to elucidate tricky metaphysical questions in philosophy but also provides a particularly promising way to think about children's causal knowledge.

As noted, much developmental research on causal reasoning has looked at children's understanding of domain-specific causal mechanisms (Bullock, Gelman &

Baillargeon et al, 1982; Leslie & Keeble, 1997; Meltzoff, 1995; Shultz, 1982; Spelke et al, 1992; Wellman et al, 1997; Woodward, 1998; Woodward, Phillips & Spelke, 1993). Although this research tradition has successfully overturned Piaget's idea that young children are "precausal" (1930), it has followed Piaget's lead in treating knowledge of distinct physical and psychological mechanisms of causal transmission as the hallmark of causal understanding.

Specifically, developmental researchers have largely accepted the idea that causal knowledge involves knowing that causes produce effects by transfer of information or energy through appropriate intervening mechanisms. In an influential monograph on children's causal reasoning, the psychologist Thomas Shultz wrote that children understand causation "primarily in terms of generative transmission" (1982). Similarly, Schlottman writes that “mechanism is part of the very definition of a cause” (2001) and Bullock, et al (1982) conclude that the idea that “causes bring about their effects by transfer of causal impetus” is “central to the psychological definition of cause-effect relations.”

Consistent with this causal mechanism or "generative transmission" approach, psychologists have suggested that even adults prefer information about plausible, domain-specific mechanisms of causal transmission to statistical and covariation information in making causal judgments (Ahn, Kalish, Medin & Gelman, 1995). Some philosophers have also adopted a transmission perspective, arguing that causal interactions are characterized by spatiotemporally continuous processes involving the exchange of energy and momentum, or the ability to transmit “a mark” (Dowe, 2000; Salmon, 1984; 1998).

However, although the generative transmission model of causation is arguably the dominant view of causal knowledge in the developmental literature, there are several respects in which this model critically fails to account for our causal intuitions. Many events that we believe are causally connected (e.g., losing track of time and being late for class; taxing cigarettes and reducing smoking) are not, at least in any obvious way, characterized by mechanisms of transmission. Second, as the philosopher Jim Woodward observes, there is no obvious reason why it should be of value to us to distinguish those events that transmit energy or information from those that do not (2003); those aspects of causality that make it of central importance to human cognition (prediction and control) do not seem to be captured by the concern with spatial and energy relations that characterize the transmission view. Furthermore, nothing in the generative transmission model distinguishes causally relevant from causally irrelevant features of transmission. Generative transmission models fail to explain why, for instance, the momentum transferred from a cue stick to a cue ball is causally relevant to the ball's movement while the blue chalk mark, transmitted at the same time and in the same manner, is not (Hitchcock, 1995, this volume).

Critically, the tendency to equate causal understanding with an understanding of mechanisms of causal transmission may pose a particular problem for the theory theory. Recent research suggests that adults cannot generate a plausible account of causal mechanisms, even in domains where they consider themselves highly knowledgeable (Rozenbilt & Keil, 2002). Keil has suggested that we suffer from an "illusion of explanatory depth" and that our causal knowledge may amount to little more than "one or two connected causal beliefs" (2003). He has argued that "calling this causal knowledge

folk 'science' seems almost a misnomer" and that "The rise of appeals to intuitive theories in many areas of cognitive science must cope with a powerful fact. People understand the workings of the world around them in far less detail than they think" (2003).

If having a theory is coextensive with having an account of causal mechanisms than Keil's suggestion is troubling, particularly since an impoverished understanding of causal mechanisms is presumably even more characteristic of young children than adults. Perhaps, children's causal reasoning is not particularly sophisticated after all.

However, the interventionist account explicit in recent philosophical work and implicit in computational models such as causal Bayes nets provides a quite different account of what it might mean to have causal knowledge. In the context of a causal model, the proposition that X causes Y (X→Y) means, all else being equal, that an intervention to change the value or probability distribution of X will change the value or probability distribution of Y. That is, the causal arrows in the graphical models are defined, not with respect to their relevance to a domain, their spatiotemporal features, or their ability to transmit energy or force, but (mirroring the way causality is understood in science) in terms of possible interventions. These interventions need not actually be realized or even feasible but they must be conceivable (see Woodward, 2003 for details). A causal relation then is defined, not in terms of its physical instantiation but in terms of the real and counterfactual interventions it supports. A theory, on this view, represents a coherent and organized set of such relations, rather than necessarily involving a set of beliefs about physical processes or mechanisms.

Recently, both statisticians and philosophers have argued that this interventionist account captures precisely what it means for a variable to be a cause (see e.g., Pearl, 2000

and Woodward, 2003). Learning algorithms based on these models support novel predictions, interventions, inferences about a range of causal structures, and inferences about unobserved causes. Arguably then, knowledge of causal mechanisms and processes of transmission may not be of central importance for at least some of what we need theories to do.

Note, moreover, that an interventionist account of causal learning is consistent with, and indeed predicts many of the findings that have been associated with the generative transmission model. In looking for instance, at children's inferences about force relationships, Shultz first taught the children what types of interventions were relevant to outcomes (e.g., that striking a tuning fork in front of an open box created a sound). Shultz then struck two tuning forks; the first failed to temporally covary with the sound (because it was positioned to the side of the box) while the second did covary with the effect (because the experimenter struck the second fork and simultaneously turned the box to face the first). Children chose the first tuning fork (with the appropriate transmission relationship) as a cause and rejected the tuning fork that merely covaried with the effect.

However, the relevant covariation information for children might not be merely the temporal covariation of the tuning fork with the effect but the covariation of interventions and outcomes; that is, children could have learned that turning the box was as critical to the effect as striking the fork. Indeed, in novel cases like this, arguably the only information that children have about processes of causal transmission is the evidence of effective patterns of intervention. Given that any causal relationship (e.g., flipping a switch and a light turning on) can be instantiated by a vast number of causal

mechanisms (many types of wires, bulbs, circuits, etc.), it may make sense that children's naïve theories should focus on the connection between interventions and outcomes rather than on the myriad mechanisms that might realize it. Indeed, one of the virtues of theories may be that they enable us to make powerful predictions *despite* our often substantial ignorance about underlying processes and mechanisms (our "trifling investment in fact").

Note that scientific theories, as well as naïve ones, often remain agnostic about processes of transmission while committing to hypothetical interventions. Newton developed his theory of gravitation without knowing any mechanism that might enable masses to attract one another; Darwin developed his theory of evolution without knowing any mechanism that might make variation in the species heritable. Thus although we might say informally that Darwin posited natural and sexual selection as “mechanisms” for evolution, we do not mean that Darwin discovered spatiotemporally continuous processes by which energy or information is transferred. Rather Darwin inferred that traits that enhance an organism’s reproductive success will be more prevalent in the population; that is, that changes to one set of variables will affect the outcome of other variables. Thus scientific theories, like naïve ones, are not necessarily derived from, or committed to, particular causal mechanisms. Rather, in identifying the causal structure – the real and hypothetical interventions the variables support -- theories help narrow the search space for the relevant physical processes.

Critically, we do not mean to suggest that substantive assumptions about spatiotemporal relations and domain-specific knowledge do not play a fundamental role in children's causal understanding. Indeed, one of the important challenges for cognitive

science is to understand how knowledge about particular physical relations in the world is integrated with evidence about interventions and patterns of covariation. In what follows we will discuss some important interactions between children's substantive causal knowledge and formal learning mechanisms. Even more critically, we do not mean that children only learn causal relations from interventions. Children may infer causal relations in myriad ways, including from spatial relations, temporal relations, patterns of covariation, and simply by being told. The claim rather, is that certain patterns of interventions and outcomes indicate causal relationships and when children infer that a relationship is causal, they commit to the idea that certain patterns of interventions and outcomes will hold.

One of the exciting features of the interventionist account of causation is that, together with theory theory, it generates an array of interesting and testable predictions about children's early learning. At a minimum, if children's causal knowledge takes the form of naïve theories and if causal knowledge is knowledge that supports interventions, children should be able to: A) use patterns of evidence to create novel interventions B) do this for any of a variety of possible causal structures; C) use evidence from interventions to infer the existence of unobserved causes; D) distinguish evidence from observation and intervention in their inferences about causal structure E) effectively weigh new evidence from interventions against prior beliefs, and F) distinguish good interventions from confounded ones.

In what follows, we will walk through this alphabet of inferences. We will discuss respects in which the causal Bayes net formalism provides a normative account of these components of theory formation and we will review evidence from our lab suggesting

that very young children are capable of this type of learning. [1]

**A) Making novel interventions**

In the absence of theories, you could safely navigate a lot of causal territory. Classical conditioning, trial-and-error learning, and hard-wired causally significant representations (of the sort that make nestlings cower when hawks fly overhead -- and arguably of the sort that is triggered by seeing one object strike and displace another, (e.g., Michotte, 1962)) are effective ways of tapping into real causal relations in the world. Each of these abilities lets us track regularities in the environment and predict some events from the occurrence of others. Some of these abilities even support effective interventions.

Like other animals, human beings seem to have innate, domain-specific causal knowledge (Spelke et al, 1992), the ability to detect statistical contingencies (Saffran, Aslin & Newport, 1996), and the ability to learn from the immediate consequences of our own actions (Rovee-Collier, 1980; Watson & Ramey, 1987). Unlike other animals however, we routinely use the contingencies and interventions we observe to design novel interventions. We routinely meet regularities with innovation.

Some of this inferential power may come from the way that human beings represent causal knowledge. Elsewhere (see Gopnik et al., 2004) we have suggested that causal Bayes nets representations provide a "causal map" of events in the world. The analogy to a spatial map is helpful because it explains both some of the advantages of the

[1] Throughout, we will assume some familiarity with the causal Bayes nets formalism (that is, we will assume that readers have already read the introduction to this book). Thus we will use terms like causal graphs, the causal Markov assumption, and conditional independence and dependence without definition.

causal Bayes net representation and some of the disadvantages of alternative ways of storing causal knowledge.

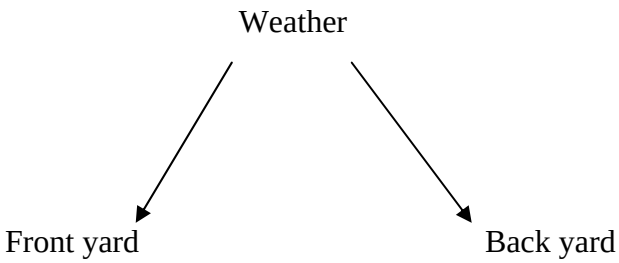
Some animals, like ants, seem to represent spatial relations egocentrically. Ants know where their nest is in relation to their own body movements but if they are scooped up and displaced even slightly, they lose their way, even in familiar terrain (Sommer & Wehner, 2004). Other animals, like mice, construct spatial maps. Once mice have explored a territory, they can always take the shortest route to a goal, no matter where they are placed initially (Tolman, 1932). Such cognitive spatial maps reveal the underlying stability of geometric relations.

Causal relations can also be represented egocentrically, in terms of the immediate outcome of one's own actions (e.g., as in operant learning). However, like an egocentric spatial representation, operant learning fails to represent the relationship of variables to one another. Operant learning restricts you to learning the immediate outcome of your own actions, and even these can only be learned by trial and error. However, if you represent causal events as they relate to one another, then -- even if you are not part of the causal structure, or even if you own relationship to the events changes -- the stability of the underlying causal structure is preserved. From such stability, may come the ability to negotiate novelty.

Causal Bayes nets provide just such a coherent, non-egocentric representation of the causal relationship among events. In a literature rife with stories about cigarette smoking, stained fingers, and lung cancer; birth control pills, thrombosis, and strokes, and prisoners, sergeants, and firing squads, almost any concept can be illustrated with a macabre example. We work with preschoolers however, so we will make use of a more

benign, indeed suburban, illustration (adapted from Pearl, 2000): Suppose you walk outside and see that the grass in your front yard is wet. You might guess that it has rained. Because you believe the weather is a common cause of the state of your front yard and your back yard, you will be able to infer that the grass in your backyard is most likely wet as well. You could represent this causal structure as the causal Bayes net in Figure 1 below, where each node is a binary variable taking either the value wet or dry.

Figure 1: A causal Bayes net




In this causal structure, the state of the front yard and the state of the back yard are dependent in probability. Knowing something about the front yard will tell you (in probability) something about the state of the back yard. That is, you can use knowledge of the causal graph and the known value of some variables in the system to predict the (otherwise unknown) value of other variables.

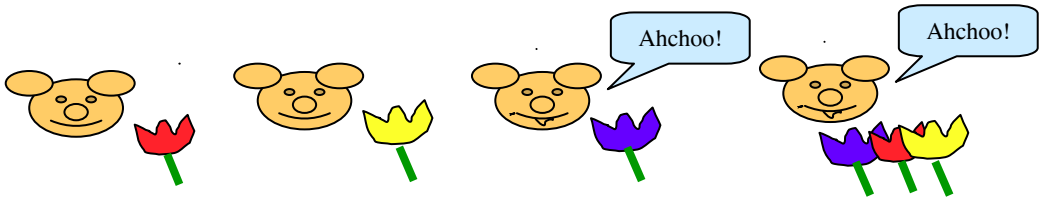
However, the critical thing about causal Bayes nets, indeed the thing that makes them causal, is that they can also support inferences about the effects of interventions. We will discuss interventions in more detail in the following section, but roughly speaking, the arrow in the graph between the weather and the front yard encodes the proposition that, all else being equal, changing the state of the weather will change the state of the front yard. Importantly, the arrow retains this meaning even though (in the real world) we can't actually intervene on the weather (short of global climate change

anyway). Knowing the causal graph lets you predict the outcome of interventions – whether or not you've ever seen them performed and indeed whether or not you could ever perform them. Thus unlike hard-wired representations or trial-by-error learning, causal graphs support genuinely novel inferences.

However, the absence of the arrow between the front yard and the back yard is also informative. Although the state of the yards are dependent in probability, there is no direct causal link between them -- all else being equal, changing the one will not change the other. Causal graphs thus represent the distinction between predictions from observation (if the front yard is dry, the back yard is probably dry as well) and predictions from intervention (wetting the front yard will not wet the back yard).

In a series of experiments, we looked at whether, consistent with the formalism, young children could use patterns of dependence and independence to make novel predictions and interventions (Gopnik, Sobel, Schulz & Glymour, 2001; Schulz & Gopnik, 2004). We showed preschoolers, for instance, that three flowers were associated with a Monkey puppet sneezing (see Figure 4 below). One flower (A) always made the Monkey sneeze, while the other flowers (B and C) only made the Monkey sneeze when flower A was also present.

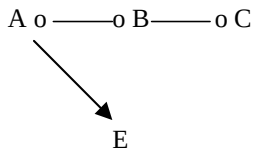
Figure 2: Evidence about three flowers




Formally, A and the effect were unconditionally dependent while B, C and the effect were independent conditional on A. Applied to this case (and assuming no

unobserved common causes) a Bayes nets learning algorithm will construct the graph in Figure 3 below.

Figure 3: Graph representing inference that flower A screens-off B and C as a cause of E.




The graph says that A causes the effect and B and C do not. (It also says that there is an undetermined causal link between A, B and C, represented by the circles and the ends of the edge connecting those variables. In fact, there is such a link, namely the experimenter, who put all three flowers in the vase together.) This structure in turn generates predictions about interventions. In particular, it implies that an intervention on A will change the value of C, but an intervention on B or C will not have this effect.

Children were asked "Can you make it so that Monkey won't sneeze?" Consistent with the prediction of the formalism, children screened-off flowers B and C and removed only flower A from the vase. Control experiments established that the inference was due to the pattern of conditional dependence and independence, not to frequency information.

One might argue however, that children have only a very limited ability to make novel and appropriate inferences. Children might, for instances, be able to use patterns of dependence to differentiate equally plausible causal candidates within a domain (i.e, the causal power of one flower vs. another). However, innate or domain-specific knowledge might restrict the range of evidence children are willing to consider in the first place. Formal inference procedures might not be able to override or change children's prior

beliefs.

However, if, consistent with the theory theory, children develop their causal understanding *from* patterns of evidence, then domain-specific judgments ought to be defeasible. Given appropriate evidence, children ought to be able to override prior knowledge and reason about truly novel events, including events that cross the boundaries of domains, and design truly novel interventions accordingly. In order to look at the extent to which children could flexibly use evidence and formal inferential procedures to make genuinely novel causal inferences, we pitted children's domain-specific knowledge against patterns of evidence.

We showed children for instance, that three causes were associated with a machine turning on. Two of the causes were domain-appropriate (buttons) and one was domain-inappropriate (talking to the machine). Talking to the machine and the machine turning on were unconditionally dependent but conditional on talking, the buttons were independent of the effect. Thus the structure was formally identical to the structure in Figure 3. We asked the children if they could turn off the machine. In a baseline condition, we provided children with no evidence and simply asked the children whether talking or pushing buttons was more likely to turn off the machine.

Consistent with past research showing that children's causal inferences respects domain boundaries, children in the baseline condition chose the domain-appropriate causes (the buttons) at ceiling. However, consistent with the predictions of the formalism, when asked to turn off the machine, 75% of the children ignored the buttons and said, "Machine, please stop." Children were able to use the pattern of conditional dependence and independence to create a new "causal map" and to generate an

appropriate, but novel, causal intervention.

In this experiment the relations between causes and effect were deterministic. Such definitive evidence might have made it particularly easy for children to override their prior knowledge. However, in another experiment (Kushnir, Gopnik & Schaefer, 2005), we tested whether children's domain-specific preference for contact in physical causal relations could be overridden in light of probabilistic evidence that physical causes could act at a distance. We showed children a toy with a colored surface and told them, "Sometimes the toy lights up." Without further instruction, we gave children a block and asked them to make the toy light up. Thirteen out of sixteen children (81%) demonstrated a strong initial assumption of contact causality, touching the block to the surface of the toy (the other three did nothing). After their intervention, we showed children four pairs of blocks. In each pair, one block activated the toy 1/3 of the time and always by contact. The other block activated the toy 2/3 of the time and always at a distance (i.e., by being held 5-6 inches above the toy). At the end of the experiment, we asked children to make the toy light up again. A significant number of children revised their original intervention and activated the toy at a distance (McNemar's test, $p<.05$). Thus children seem to be able to revise their domain-specific knowledge and create novel interventions, even when given only stochastic evidence for new causal relations.

If children's causal reasoning were constrained by innate representations or informationally-encapsulated modules, such flexibility and sensitivity to evidence would be surprising. However, it is less surprising from a theory theory perspective. The ability to overturn prior knowledge and learn something genuinely new is one of the chief virtues of scientific inquiry. It may also be one of the hallmarks of childhood.

## B) Learning a wide range of causal structures

If you were a Martian reading much of the classic literature on human causal reasoning, you might assume that Earth was a relatively simple place. The stakes are sometimes high (Does camouflage protect tanks from being blown-up? Does gender affect college admissions? Does medication cause headaches? Baker et al., 1989; Bickel et al., 1975; Hammel, & O'Connell, 1975; Novick & Cheng, 2004) but the questions, at least, are straightforward: Given a particular set of evidence, is C a cause of E?

Many theories have tried to explain how people answer this question. Accounts ranging from the associative learning accounts discussed above to Patricia Cheng's elegant power theory of probabilistic contrast (Cheng, 1997; Novick & Cheng, 2004), have looked at how people might estimate the relative strength (or, uniquely in Novick & Cheng, 2004, the conjunctive strength) of variables to produce an outcome.

However, both the question and the ways we might answer it assume that variables in the world are already identified as (potential) "causes" or as "effects". A Martian might reasonably wonder whether events on Earth come with labels. The question does not ask, and the theories do not answer, how we might distinguish causes from effects in the first place. Put another way, both associative learning accounts and the power PC account aim to explain how people distinguish the *strength* of different causal variables. They do not explain how people make judgments about causal *structure*.

Sometimes of course, events in the world *are* essentially "labeled" by the information around them. Spatial cues, combined with prior knowledge about plausible causal mechanisms, may identify some variables as potential causes and others as effects. In other cases (not coincidentally including camouflage and explosions, gender and

college admissions, medicine and headaches temporal priority makes the distinction transparent (Lagnado et al, this volume).

However, spatiotemporal cues are not always available in the input. If cause and effect occur at nearly the same time (the dog barks and the cat runs) or if you walk in on the middle of a scene (brother is sulking and sister is mad) there may be no way to know "who started it". Moreover, even when temporal cues are present, they may be misleading. A naïve learner who sees Mom search under the bed and then exclaim with joy upon finding her car keys might be justified in concluding that searching caused Mom to want her keys rather than that desire motivated the search.

More critically, any theory (naïve or scientific) requires knowing something more than the set of binary relations (does X cause Y?) that obtain between events. A prerequisite to theory formation must be the ability, not just to distinguish the strength of causal variables, but to organize variables within a causal structure. Indeed, part of what differentiates a theory from an empirical generalization is that within a theory, causal relations are coherent and mutually reinforcing.

The causal Bayes net formalism provides a way to represent and learn complex, coherent causal structures without prior knowledge about whether variables are causes or effects. Although the formalism *can* incorporate background information from prior knowledge, substantive cues, and temporal order (see section E of this chapter), the direction of causal arrows can also be derived directly from the patterns of conditional dependence and independence in the data. Some structures can be distinguished by observation only; others require a combination of observation and interventions.

Suppose for instance, that you see three correlated events and are trying to decide whether A and B cause C or whether C causes A and B. If the causal structure is a common effect (A→C←B) you are more likely to see A and C co-occur and B and C co-occur than to see A and B co-occur. However, if the structure is a common cause (A←C→B), you are likely to see all three variables co-occur. B will be independent of A conditional on C in the common cause case but not the common effects case. These structures can be distinguished just by observation.

The situation is more complex if you are trying to distinguish other structures. For example, suppose you are trying to distinguish the common cause structure (A←C→B) from the causal chain (A→C→B). In the common cause structure, if C occurs exogenously, it will activate both A and B and you will tend to see all three variables together. Similarly, in the chain, if A occurs exogenously it will activate C which will activate B and again, you are likely to see all three variables co-occur. In both cases B is independent of A conditional on C. Such "Markov-equivalent" structures are indistinguishable under observation. However, these structures can be distinguished by intervention. If you intervene to make C happen, you will increase the probability of seeing A and B if the structure is a common cause (A←C→B) but will have no impact on the probability of observing A if the structure is a chain (A→C→B) (See Steyvers, Tenenbaum, Wagenmakers & Blum, 2003 for discussion and evidence that adults are sensitive to these distinctions). Given a combination of evidence from observation and intervention, the causal Bayes net formalism allows for learning the structure even of very complex, multi-variable systems.

Within the formalism, interventions are treated as variables with special features. Specifically, they must be exogenous (that is they must not be influenced by any other causal factors in the graph) and they must fix the value or probability distribution of the variables of interest. After an intervention, the value of the intervened-upon variable is entirely determined by the intervention and not by any pre-existing causes (see Figure 3 below). Thus interventions on a Causal Bayes net break arrows *into* the variables of interest, performing what Judea Pearl vividly described as "graph surgery" (2000). We can then look at the "post-surgical" graph (after the intervention has taken place) and figure out what has happened to the other variables in the graph.

Figure 4
*4a*) A causal chain.

X→Y→Z

*4b*) A causal chain after "graph surgery"; the intervention on Y breaks the arrow between X and Y.

X I→Y→Z

There are several different ways of formally capturing these relations between interventions, dependencies and causal arrows (see Pearl, 2000; Spirtes, Glymour, & Scheines, 1993; Woodward, 2003). One way to do this is in terms of what we have called the conditional intervention principle. The conditional intervention principle can be formally stated as follows: for a set of variables in a causal graph, A directly causes B (that is, A→B) if and only if: 1) there is some intervention that fixes the values of all other variables in the graph and results in B having a particular probability distribution, pr (Y); such that 2) there is another intervention that changes the value of A and 3) changes the probability distribution of B from pr (B) to pr' (B) but 4) does not influence

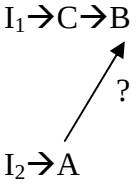
B other than through A and 5) does not undo the fixed value of the other variables in the graph (Gopnik et al., 2004).

Although this principle may sound complex, it is simply a formal statement of the sort of intuitions about intervention and causation that underlie experimental design. In an experiment, if you want to find out the causal relationship between two variables, you intervene to hold all other variables constant, and then you intervene to manipulate the value of the variable of interest. If, for instance, you want to know the causal relationship between A and B (represented by an arrow with a question mark in Figure 3a below) you can perform one intervention ($I_1$ in Figure 5a) to hold all other potential causes of B constant and another intervention to change the value of A ($I_2$ in Figure 5a). If the value (or probability distribution) of B changes, you can conclude that A causes B.

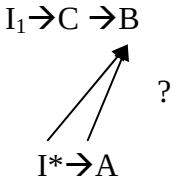
Note also that the conditional intervention principle rules out confounded interventions. Line 4 of the conditional intervention principle eliminates the graph in 5b (because the intervention on A cannot influence B except through A) and line 5 rules out the confounded graph in 5c (because interventions cannot change the fixed value of any other variable in the graph).

Figure 5. Graphs illustrating the conditional intervention principle.

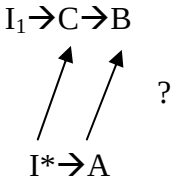
*5a*) $I_1$, fixes the value of other causes of B (clause 1 of the conditional intervention principle). $I_2$, changes the value of A (clause 2 of the conditional intervention principle).



*5b*) I* is ruled out by clause 4 of the intervention principle because the intervention affects the value of B directly.




*5c*) I* is ruled out by clause 5 of the intervention principle because the intervention affects other causes of B.




Motivated by causal Bayes net theory (see also Lagnado et al, Hagmeyer et al, Rehder, Tenenbaum & Griffiths, this volume), researchers have recently shown that adults can make appropriate inferences about a wide range of causal structures beyond simple cause-effect parings. Importantly, the evidence suggests that causal strength learning (and subsequent inferences) can and does take place in the context of complex causal models. For example, Waldman (2000, 2001) has shown that adults are sensitive to the direction of causal arrows when learning and reasoning about causal strength relations – that is, they make the distinction between "predictive" and "diagnostic" inferences – a fact that cannot be predicted based on associative learning mechanisms alone. Other studies (Waldman & Hagmayer, 2005; Lagnado et al, this volume; Sloman & Lagnado, 2005**)** have shown that, given causal models, adults can make inferences

about the effects of hypothetical interventions as well. Thus, psychologically, causal strength judgments do not take place outside of the context of causal structures.

All this should satisfy a Martian that adult humans can make appropriate predictions about observations and interventions in a broader causal context. But of course, adult humans, particularly the university undergraduates tested in these studies, have extensive experience and often quite explicit tuition in causal inference. Moreover, for the most part these studies have focused on making inferences about evidence given knowledge of a particular structure, rather than learning structure from evidence. These studies do not tell us whether this sort of causal learning is part of a more fundamental human learning mechanism, and in particular whether it might be responsible for the impressive learning we see in very young children. Conversely, the studies of children we have just described all presented them with the classical problem of inferring which cause was responsible for a particular effect – which blicket set off the detector, which flower made monkey sneeze. In principal, these results might be explained by variations of earlier theories such as associationism or causal power theory. Nor have studies so far tested explicitly whether adults or children can use the conditional intervention principle to make inferences about complex causal structures, as the Bayes net formalism would suggest. In the absence of distinguishing spatiotemporal information, can children use evidence from observations and interventions to learn the structure of causal chains, common effects, common causes and causal conjunctions?

To find out, we introduced preschool children (mean age 4;6) to a gear toy. Children saw that when a switch was flipped, two gears, A and B, spun simultaneously. There were four possibilities: (i) the switch activated gear A and A made B go (ii) the

switch activated gear B and B made A go (iii) the switch activated each gear independently or (iv) the switch activated the gears but neither gear would spin without the other. Note that these structures are indistinguishable under observation; no matter which structure obtains, when you flip the switch, both gears will spin together.

The structures are however, distinguishable under intervention. If for instance, you remove gear B, flip the switch on and gear A spins, you can eliminate structures (ii) and (iv). If you replace gear B, remove gear A, flip the switch on and gear B fails to spin, you can eliminate structure (iii) and infer that structure (i) is correct. This type of inference is a direct application of the conditional intervention principle. Controlling for other causes of A (the state of the switch), an intervention on A changes the value of B (when the switch is on and A is present, B spins; when A is absent B does not) whereas controlling for other causes of B, an intervention on B does not change the value of A. You should conclude that structure (i) is correct and A→B. Because the patterns of evidence under intervention are unique to each structure, the correct structure can be determined from the data that result from interventions.

Over a series of experiments we found that, consistent with the formalism, four-and-a-half-year-olds were able to learn the correct causal structure, represented by a simple picture, from the type of evidence described above. Children were equally good at learning all four structures (the two chains, the common effect and the conjunction). In each case, when children were presented with the appropriate evidence, they chose the correct structure significantly more often than any of the other structures. Control experiments suggested that children's judgments were not based on substantive cues or prior knowledge about gears. Additionally, consistent with the data reported in section A

of this chapter, children were able to use their knowledge of the causal structure to make novel predictions. Children who had never seen gears A and B on the toy but were told the structure (e.g., that A spun B) were able to predict the evidence that would result from interventions (e.g., that when the switch was on and A was on the toy by itself, A would spin, but that when B was on by itself, B would not). Again, children were equally good at predicting the outcomes of interventions for all four structures. (Schulz, 2003; Schulz, Gopnik & Glymour, in submission).

These experiments are particularly noteworthy because they were explicitly inspired by the Bayes net formalism and are not explicable by any other existing theory of causal learning. The physical and mechanical features of the gears were identical in all cases and the associations and covariations between the gears were also held constant. The complex pattern of relations between interventions and observations allowed children to learn complex causal structure – in just the way the formalism would suggest.

In their everyday life children intervene widely on the world and see a wide range of interventions performed by others. At least in simple, generative, deterministic cases, preschool children seem to be able to infer a range of different causal structures from patterns of evidence, and to predict patterns of evidence from knowledge of causal structure. Even very young children seem to rely on some of the same formal principles of causal inference that underlie scientific discovery. Such mechanisms may help children to develop intuitive theories of the world around them.

### C) Inferring the existence of unobserved causes

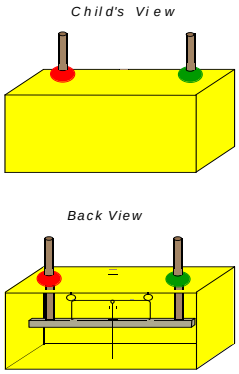
One of the critical respects in which science sometimes brings us genuinely new insight is by invoking unobserved causes to explain events. However, unobserved causes

are not the exclusive provenance of scientific theories. Children's naïve physics relies upon unobservable forces, children's naïve psychology upon unobservable mental states, and children's concept of natural kinds upon unobservable essences (e.g., Bullock et al, 1982). It is thus perhaps surprising that most psychological accounts of causal reasoning (Shanks & Dickinson, 1987; Cheng, 1997) relegate unobserved causes to a background condition.

We have already discussed respects in which the causal Bayes net formalism supports inferences about the unknown *value* of some variables from the known value of others. However, in some cases the formalism supports inferences about the *existence* of variables themselves. In particular, if the known values in the graph generate patterns of conditional dependence and independence that appear to violate the causal Markov assumption, the formalism infers the existence of an unobserved cause.

In a series of experiments, participants (both adults and children) were introduced to a "stickball machine" (see Figure 3). The two stickballs could move up and down (either simultaneously or independently) without any visible intervention (because they could be manipulated from behind the machine). The experimenter could also visibly intervene on a stickball by pulling up on the stick. This might cause – or fail to cause -- the other stickball to move.

Figure 3: The Stickball Machine




We looked at whether, consistent with the causal Markov assumption, adults and kindergarteners could use interventions and the pattern of outcomes to infer the existence of an unobserved common cause. In these studies, participants saw that the movement of the two stickballs was correlated in probability. They then saw that an intervention on stickball A (pulling on A) failed to move B and that an intervention on B failed to move A. On comparison trials, participants were given evidence consistent with A→B (e.g., they saw that pulling on B failed to move A, but they did not see an intervention on A).

If the movements of A and B are probabilistically dependent but intervening to "do A" fails to increase the probability of B moving and intervening to "do B" fails to increase the probability of A moving, the causal Markov assumption can be preserved only by inferring the existence of an unobserved common cause of A and B (i.e., that the true causal structure is: A←U→B). This structure predicts the observed evidence: A and B are unconditionally dependent in probability but an intervention on either A or B breaks the dependence.

Consistent with the formalism, both adults and children inferred the existence of an unobserved common cause when interventions on either stickball failed to correlate with the movement of the other. Adults drew the appropriate

graph (A←U→B); children inferred that "something else" (besides either of the stickballs) was making the stickballs move (Kushnir, Gopnik, Schulz & Danks, 2003; Schaefer & Gopnik, 2003). Importantly, participants only postulated an unobserved common cause when no other graph was consistent with the observed pattern of dependencies. The causal Bayes net formalism thus provides a mechanism by which evidence about observed variables can lead to inferences about the existence of unobserved variables. Processes like these might help explain how both children and scientists bring new theoretical entities into the world.

**D) Distinguishing evidence from observations and interventions**

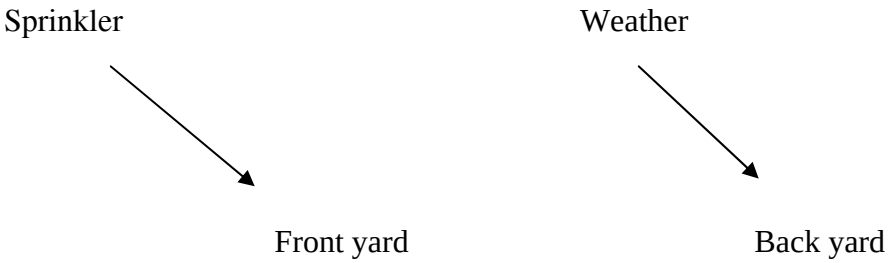
At the core of the theory theory is the idea that children learn causal structure from evidence. There are two ways we can get (firsthand) evidence about an event: we can see the event happen or we can make the event happen. Importantly as we have implied in the previous sections, these two ways of getting data -- seeing and doing -- can lead to radically different conclusions, even when the evidence itself is otherwise identical. What you can learn depends, not only on what you already know, but on how you know it.

In section A) we discussed a simple causal graph in which the weather was a common cause of the state of the front yard and the back (F←W→B). We noted that using this graph, you could predict the state of the backyard from the state of the front yard.

Suppose however, that you buy a sprinkler for your front yard and set it to go off every morning at six. Setting the sprinkler cuts the arrow between the weather and the

front yard and breaks the dependence between the front yard and the back. The altered graph is shown in Figure X below.

Figure X: A causal Bayes net with a sprinkler




If the graph is as depicted in Figure X, when you look outside and see that the grass in your front yard is wet, you will not be able to infer that the grass in your backyard is also wet. Evidence that was informative under observation is uninformative under this intervention.[2]

One of the strengths of the causal Bayes net formalism is that it supports accurate inferences whether the evidence comes from observations, interventions, or combinations of the two. Because the causal graph under intervention is different from the graph under observation, the same evidence should lead to different inferences.

The theory theory implies that young children should be sophisticated causal reasoners. Are children also sensitive to the distinction between evidence from observations and evidence from interventions and do they modify their inferences

[2] We kept the example deliberately simple. Of course if the state of the grass were measured as a continuous variable (how wet is it?) rather than as a binary one (is it wet or dry?) then you might observe that the front yard was *wetter* when it had rained than when it had not. In that case, the intervention to set the sprinkler would not break the arrow between the weather and the grass and knowing something about the front yard would still tell you something about the back. The arrow would be similarly preserved if you invested in an expensive sprinkler that only turned on when it hadn't rained – which might be a better choice for your lawn but (because the state of the sprinkler is no longer exogenous to the graph) a bad example of an intervention.

accordingly? Note that, such sensitivity is not predicted by all models of causal reasoning. Accounts of causal reasoning that use the strength of the association between two variables as indicative of the probabilistic strength of the causal connection between them (see e.g., Dickinson, Shanks & Evenden, 1984; Shanks, 1985; Shanks & Dickinson, 1987; Wasserman, Elek, Chatlosh, & Baker, 1993) are indifferent to whether the association is due to intervention or observation. Because of this, the predictions made by causal variants of the Rescorla-Wagner equation and the causal Bayes net formalism sometimes differ.

In a series of experiments (designed primarily to look at children's ability to distinguish common cause structures from causal chains) we looked at whether children's conclusions changed depending on whether they observed the relevant evidence with or without an intervention. Children were introduced to the "stickball machine" described in Section C. Children were told "Some stickballs are special. Special stickballs almost always make other stickballs move." Children were taught that one stickball might be special, both stickballs might be special, or neither stickball might be special.

In the test condition, children saw the stickballs move up and down simultaneously (without an intervention) three times. The experimenter then visibly intervened by pulling on the top of one stickball; the other stickball failed to move. In the control condition, the experimenter intervened by pulling on one stickball and both stickballs moved simultaneously three times. The experimenter then pulled on the stickball a fourth time and the other stickball failed to move. At the end of the trials, the experimenter pointed to each stickball and asked "Is this stickball special?"

In the test condition, there is a correlation between *seeing* stickball Y move and seeing stickball X move. However, intervening to move Y breaks the dependence. From a causal Bayes net perspective, this pattern of evidence is consistent with the graph X→Y but not with the graph Y→X. Children should say that X is special but deny that Y is special. In the control condition, *intervening* on Y and seeing X move are probabilistically dependent throughout. This is consistent with Y→X but not X→Y; children should say that Y is special and X is not.

Note however, that from an associative learning perspective, the strength of association between the stickballs is the same in both conditions. The movement of stickball Y is associated with the movement of stickball X every time but one. If children are reasoning associatively, then in both conditions they should say that Y is special.

The children (four-and-a-half-year-olds) distinguished between evidence from observations and interventions and reasoned, not as predicted by associative learning models, but as predicted by the causal Bayes net formalism. That is, children were significantly more likely to affirm that X was special and deny that Y was special in the test condition than in the control condition, and significantly more likely to affirm that Y was special and deny that X was special in the control condition than in the test (Schulz, 2001; Gopnik, et al., 2004).

Similarly, in the unobserved cause studies discussed in Section D, we reported that participants saw that *intervening* to move stickball X failed to move stickball Y and *intervening* to move Y failed to move X. In control conditions however, participants saw X move by itself and Y move by itself, but this time the stickballs moved without visible intervention – the experimenter simply *pointed* at X when it moved by itself and then

*pointed* at Y while it moved by itself. Consistent with the predictions of the formalism, participants distinguished between the two conditions and only inferred the existence of an unobserved common cause of X and Y (X←U→Y) in the intervention condition. (In the observation condition, they inferred the existence of two independent unobserved causes: U1→X and U2→Y.)

Pearl writes that "Scientific activity, as we know it, consists of two basic components: Observations and interventions. The combination of the two is what we call a laboratory . . ." (2000). Although making inferences about stickball machines may seem a far cry from scientific inquiry, the ability to distinguish evidence from observations and interventions is fundamental to both. Sensitivity to the different role played by these "basic components" may help support children's ability to learn the causal structure of events in the world.

**E) Weighing new evidence against old beliefs**

A few pages back, we reported that preschoolers ignored a machine's buttons and asked a machine to stop after seeing – once – that talking and the toy activating were unconditionally dependent. We reported this as partial proof of the cleverness of four-year-olds. This might worry you. This might also worry our Institutional Review Board. Are preschoolers unreasonably impressionable? Surely it's not that clever to override the whole of naïve physics on the evidence of a single trial. Surely -- even in Berkeley -- we don't want children going around talking to machines. Learning flexibly from evidence is all very well, but can causal Bayes nets run amok?

Well, no – at least not in this respect. Causal Bayes net representations can be inferred by a variety of different learning algorithms discussed constraint-based and

Bayesian learning algorithms (see Gopnik et al, 2004 for discussion). Both of these algorithms can take prior knowledge into account. Constraint-based algorithms test pairs and triads of variables for independence and conditional independence. By adjusting the significance level of the statistical test used to determine independence, constraint-based methods ensure that variables likely to be independent based on prior knowledge (e.g., talking and a machine activating) are subject to less rigorous tests of independence than variables that, given prior knowledge, are less likely to be independent.

A somewhat more elegant approach is adopted by Bayesian causal learning methods. Bayesian algorithms assign all the possible causal hypotheses (the causal graphs) a prior probability. This probability is then updated given the actual data (by application of Bayes theorem). The posterior probability of each causal graph is evaluated to see which model best fits the data. Thus it will take more evidence to support an initially unlikely causal hypothesis than an initially probable one.

Several studies show that under conditions of uncertainty, people do take current evidence and prior knowledge into account as predicted by Bayesian learning algorithms (Griffiths, 2004; Griffiths & Tenenbaum, 2001; Tenenbaum & Griffiths, 2003, Tenenbaum and Griffiths this volume). In one study for instance, adults were taught that "super pencils" would activate a "superlead" detector. During a training period, adults were taught that super pencils were either rare (2 of 12 pencils activated the detector) or common (10 of 12 activated the detector). Two (previously untested) pencils were then placed on the detector and adults saw that both pencils (A and B) together activated the detector and also that A by itself activated the detector. The adults were asked to estimate the likelihood that B by itself would activate the machine.

As predicted by the Bayesian learning algorithms, (but not as predicted by associative learning accounts) prior knowledge about the prevalence of super pencils affected people's causal judgments. Despite seeing identical evidence about B in both conditions, participants believed B was much more likely to activate the machine in the common condition than in the rare condition (Tenenbaum & Griffiths, 2003). Other studies showed that four-year-old children could make similar judgments. Taught either that blickets were rare or common, and shown the "backwards blocking" condition described above, children inferred that B was a blicket when blickets were common and that B was not a blicket when blickets were rare (Sobel, Tenenbaum, & Gopnik, 2004; Sobel & Kirkham, this volume, Tenenbaum & Griffiths, this volume).

So if preschool children take prior knowledge into account when making causal judgments, why did children in the talking machine experiment violate their knowledge about domain-appropriate causes on the evidence of a single trial? Note that in the cross-domains experiment, children were given deterministic data: Buttons and the machine turning on were *always* independent conditional on talking; talking and the machine turning on were *always* unconditionally dependent. When evidence is deterministic, you don't need statistical tests to determine independence and whatever the prior probability of the hypothesis, the posterior probability is 100%. Given the deterministic evidence, children's inferences were identical to those that would be made by the formalism.

Importantly however, in a more ambiguous scenario, children did take prior knowledge into account. We replicated the machine/talking experiment with a new group of children and then tested the children on a "transfer condition" with a novel toy, a

novel speech act, and two novel switches. In the transfer condition, children received no evidence about the novel stimuli; we simply asked the children how they would activate the novel toy: by talking to it or by flipping the switches. In the test condition, the children talked to the machine, just as in the previous study. However, in the transfer condition, despite the similarity of the stimuli, the children largely reverted back to their prior knowledge: 75% of the children chose the switches (the domain-appropriate cause).

Equally importantly however, the prior exposure to the domain-inappropriate evidence did affect children's causal judgments. Children were significantly more likely to choose the domain-inappropriate cause in the transfer condition than in the previous baseline condition (i.e., where they had no evidence whatsoever about domain-inappropriate causes). The recent exposure to counter-intuitive evidence affected how children extended their causal inferences. Similarly, as discussed in Section A, we found that many children would override their preference for contact in physical causal relations in light of probabilistic evidence for action-at-a-distance. Thus the combination of prior knowledge and formal inference procedures seems to allow for learning that is both conservative and innovative.

This tension between conservativism and innovation is consistent with a theory theory approach to conceptual development and is also a salient feature of adult scientific inquiry. Surprising evidence is often questioned or dismissed before it is taken seriously enough to establish the theories that will, in turn, make the evidence predictable. As William James (perhaps apocryphally) is said to have quipped: "When a thing is new, people say: 'It is not true.' Later, when its truth becomes obvious, they say: 'It is not

important.' Finally, when its importance cannot be denied they say: 'Anyway, it is not new.'"

As scientists, we may kvetch about the tendency of prior beliefs to squelch innovation, however, as an extension of the inferential procedures used in childhood, the advantages of carefully weighing new evidence against old is clear. If children's learning were too flexible -- if it were, for instance wholly dictated by the most recent evidence observed -- then children would be subject to endless error. Children live in a noisy world and might easily be exposed to misleading data. If, on the other hand, innate or prior knowledge acted as a strong constraint on children's causal learning, then errors made early in development would be irreparable. Children would be intransigent in the fact of corrective evidence and helpless in genuinely novel environments.

Although science has a reputation for objectivity, one of the advantages of having a theory (naïve or scientific) is precisely that all evidence is *not* treated equally. By limiting the evidence to which we attend, or which we take seriously, theories explain in part why science can get so much inferential power out of a "trifling investment in fact". Formal inference procedures, able to take into account both prior knowledge and new evidence, may provide just the sort of learning mechanism that allows children's causal theories to be both stable and defeasible.

**F) Distinguishing good interventions from confounded ones**

People who get exercised by the concept of child as scientist frequently point out what is indisputably the case: that children, unlike scientists, do not go around designing controlled experiments to test their theories. Moreover, when children do try to design experiments (i.e., because a teacher or a researcher asks them to) they perform poorly.

Children tend to intervene on many variables at once, change interventions between conditions, and then draw all the wrong conclusions. Adults (and often scientists!) do little better (Kuhn, 1989; Kuhn, Amsel & O'Laughlin, 1988; Masnick & Klahr, 2003).

However, designing an experiment requires metacognition. To design an appropriate intervention, you have to know what makes an intervention appropriate. Learning from interventions does not require metacognition. You may have no idea what makes one intervention better than another and still be able to draw correct conclusions from the patterns of evidence that result.

In the previous sections, we provided evidence suggesting that when children are given good evidence, they draw normative causal conclusions. What happens, however, when children are given bad evidence? Are there conditions under which children realize that interventions are confounded? Does confounding change the types of inferences children make?

The conditional intervention principle defined an intervention so as to rule out instances of confounding: an intervention on X should be exogenous, should break all the arrows into X, and should not influence any other variable in the graph except through X. In the test condition of the gear toy experiment, we showed children evidence consistent with the conditional intervention principle and children were able to learn the relationship of the gears to one another.

In the control condition however, we concealed the state of the switch. Thus, just as in the test condition, children saw for instance, that gear A spun when B was removed but gear B failed to spin when gear A was removed. However, with the switch hidden, the children couldn't know whether B failed to spin because gear A was removed or

because the experimenter failed to flip on the switch. That is, there was no way to know whether the intervention to remove gear A broke all the arrows into B or not. Although the movement of the gears was the same in both conditions, children in the control conditions responded at chance and – anecdotally – tried to look behind the machine to determine whether the switch was on or off!

In a different set of studies, we looked at children's sensitivity to probabilistic causes and the role played by their own interventions. In an Observation condition, children saw an experimenter place a block on a toy three times in a row. The children saw that one block made the toy light up 2 out of 3 times while another block made the toy light up only 1 out of 3 times. Children were told that each block had "special stuff" inside and were asked which block has more "special stuff". The children distinguished the 2/3 probability from the 1/3 and said that the 2/3 block had more "special stuff".

The Intervention condition was identical except that children were allowed to intervene on the block on the third trial. For the 2/3 block, children saw the block light up the toy twice, but when they tried the block, it failed to light up. For the 1/3 block, children saw the block fail to light up the toy twice, but when they tried the block, the toy did light up. In this condition, children said that the 1/3 block had more "special stuff". Children seemed to prefer making inferences based on their own interventions.

Critically however, the children were also tested in a Confounding control condition. In the control condition, children saw exactly the same evidence as in the test condition, however this time when the child the intervened, the experimenter simultaneously pushed a button "to make the toy light up". The child's "intervention" was thus no longer a real intervention – it did not break other arrows (like the

experimenter pushing the button) into the effect. When the children's own interventions were confounded in this way, they did not express a preference for their own interventions; the children returned to judging the blocks on the basis of the probabilities (Kushnir, 2003; Kushnir & Gopnik, in press).

These findings suggest that although children may not be able to design controlled experiments, they do, at least in certain cases, recognize instances of confounding. Children seem to be sensitive to some of the fundamental features of experimental design and make different inferences when causal manipulations are consistent with the conditional intervention principle than when they are not.

Still, we might ask how, in the absence of controlled experiments, children are able to learn so much from interventions. We rely on experimental design heavily in science; how can children learn so much in its absence? Why aren't children constantly running into confounded interventions and drawing inaccurate causal conclusions?

One possibility is that the very fact of being a child might serve children well. Children are notorious for being impulsive (they get into a lot of things) and perseverative (they get into the same things over and over again). Cast in a more positive light, children tend to intervene a lot and they tend to replicate their interventions. Children's very immaturity (and in particular, the protracted development of their prefrontal cortex which (in adults) seems to inhibit impulsivity (e.g., Casey, Giedd & Thomas, 2000; Chao & Knight, 1998) and prevent perseveration (e.g., Goel & Graffman, 1995) may support causal learning.

How might immaturity and noise substitute for controlled experimental design? Note that to infer that X causes Y, you don't necessarily have to hold other causes of Y

constant. You can also randomize other causes of Y. Children's tendency to intervene in many different contexts and their tendency to replicate their actions might be advantageous. Other causes of Y (whatever Y is) might exist, but children's own actions are unlikely to always coincide with those causes. Certainly, children may occasionally leap to the wrong causal conclusion from bad evidence. Wu and Cheng (1999) for instance, cite a childhood anecdote in which one of the authors dropped a vase at the same time that a power outage occurred and thus blamed herself for the blackout. However, such anecdotes are funny in part because they are rare. In general, children's own actions may be a trustworthy foundation for their causal inferences and naïve theories.

## Conclusion

In many respects, the causal Bayes net formalism seems to provide a learning mechanism that captures the dynamic nature of theories – and in many respects, children's learning seems to be commensurate with the predictions made by the formalism. However, the causal Bayes net formalism may not tell the whole story. In particular, the formalism may not entirely satisfy Mark Twain. How we get such "wholesale returns of conjecture out of a trifling investment in fact" remains something of a mystery.

Causal Bayes net algorithms were developed for use in procedures like data mining, where evidence is plentiful but the causal relationships are obscure. Constraint-based search methods thus rely upon the evidence of many trials or assume the available data is representative of a larger sample. Bayesian learning algorithms rely either upon an abundance of data or an abundance of prior knowledge.

In our experiments by contrast, evidence was scarce. Children made causal inferences from a minimal amount of data, often using only the evidence of a single trial. As Tenenbaum and Griffiths (2003) note, in "Many cases . . . causal inference follow(s) from just one or a few observations, where there isn't even enough data to reliably infer correlation!"

Note however, that the causal Bayes net formalism was also developed to infer causal structure from noisy, probabilistic data in contexts where interventions were impossible (e.g., in epidemiological studies). By contrast, in all of our studies, children observed or performed interventions, and in most cases the evidence they saw was deterministic. Such contexts (when interventions are possible and determinism is assumed) may be plentiful in everyday life, and within such contexts, children may not need the full apparatus of the causal Bayes net learning algorithms. Children may be able to represent structure as a causal Bayes net, and may use some of the same principles about the relationship between evidence and structure, without requiring the full power of the learning algorithms (see Richardson et al, this volume). Thus the causal Bayes net formalism may be "too big" for what children need to accomplish.

Alternatively, causal Bayes nets formalism may be "too small". The algorithms may miss a level of abstraction (what Tenenbaum & Niyogi (2003) & Tenenbaum & Griffiths (this volume) call a "causal grammar") that encompasses higher order causal laws that are assumed but never explicitly presented to the children (for instance, that blocks activate detectors and detectors don't activate blocks). Children may be successful at learning causal relationships from a few observations (in our lab and in the world) because they are already bringing a rich theoretical structure to bear upon the inferential

tasks. Thus the causal Bayes net algorithms may allow children to learn structure from minimal data only when they are embedded within higher-order causal theories (see Tenenbaum & Griffiths, 2003; Tenenbaum & Niyogi, 2003, and Tenenbaum & Griffiths chapter this volume).

Critically, however, this account may only move the problem of causal inference back a step. Knowledge of higher-order causal laws might support children's ability to learn particular causal relations. However, somehow children must also learn the higher-order causal laws – and it seems tempting to assume that children infer higher-order causal laws from particular causal relations. One of the challenges for future research is to determine whether such circles can be benign rather than vicious. In principle, children might be able to bootstrap an abstract causal grammar from clear evidence for particular causal relationships, and then use the higher-order theory to handle more complex or ambiguous evidence for particular causal relations.

However, even if (as we expect) the causal Bayes net formalism does not end up being "just right", it more than any other current computational account, suggests a learning mechanism that does justice to much of the breadth and depth of children's naïve theories. In supporting novel predictions, novel interventions, structure learning, inferences about unobserved causes, distinctions between observations and interventions, and the criteria for a "good" intervention, the causal Bayes net formalism captures much that is critical about a theory. Our hope is that children's ability to engage in theory formation and theory change might similarly set the standard for future computational accounts of learning.

If you are persuaded by little else by this chapter, we hope we have at least convinced you of the value of interdisciplinary work. Research in computer science, artificial intelligence, and philosophy has suggested some of the fundamental assumptions that might underlie the development of children's naïve theories. Work in developmental psychology has demonstrated that young children are able to learn the causal structure of events with remarkable speed and accuracy. We hope that investigators in all these areas will continue to find causal learning, in both children and science, fascinating for years to come.


Ahn, W., Kalish, C. W., Medin, D. L., & Gelman, S. A. (1995). The role of covariation versus mechanism information in causal attribution. *Cognition, 57*(3), 299-352.

Baillargeon, R., Kotovsky, L., & Needham, A. (1995). The acquisition of physical knowledge in infancy. In D. Sperber & D. Premack (Eds.), *Causal cognition: A multidisciplinary debate. Symposia of the Fyssen Foundation; Fyssen Symposium, 6th Jan 1993, Pavillon Henri IV, St-Germain-en-Laye, France* (pp. 79-115). New York, NY, US: Clarendon Press/Oxford University Press.

Baker, A., Mercier, P., Valee-Tourangeau, F., Frank, R., & Maria, P. (1993). Selective associations and causality judgments: Presence of a strong causal factor may reduce judgments of a weaker one. *Journal of Experimental Psychology: Learning, Memory and Cognition, 19*, 414-432.

Bickel, P. J., Hammel, E. A., & O'Connell, J. W. (1975). Sex bias in graduate admissions: Data from Berkeley. *Science, 187*, 389-404.

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.

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

Casey, B.J., Giedd, J.N., and Thomas, K.M. (2000).Structural and functional brain development and its relation to cognitive development.*Biological Psychology, 54, 241-257.*

Chao, L. L., & Knight, R. T. (1998). Contribution of human prefrontal cortex to delay performance. *Journal of Cognitive Neuroscience, 10*(2), 167-177.

Cheng, P. W. (1997). From covariation to causation: A causal power theory. *Psychological Review, 104*(2), 367-405.

Dickinson, A., Shanks. D. R., & Evendon, J. (1984). Judgment of act-outcome contingency: The role of selective attribution. *Quarterly Journal of Experimental Psychology, 36,* 29-50.

Dowe, P. (2000). *Physical causation*. New York: Cambridge University 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(1)[243]*, v-96.

Gelman, S. A., & Wellman, H. M. (1991). Insides and essence: Early understandings of the non-obvious. *Cognition, 38*(3), 213-244.

Goel, V., & Grafman, J. (1995). Are the frontal lobes implicated in "planning" functions? Interpreting data from the Tower of Hanoi. *Neuropsychologia, 33*(5), 623-642.

Gopnik, A., Glymour, C., Sobel, D. M., Schulz, L., Kushnir, T., & Danks, D. (2004). A theory of causal learning in children: Causal maps and Bayes nets. *Psychological Review, 111*, 1-31.

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

Gopnik, A., Sobel, D. M., Schulz, L. E., & 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*(5), 620-629.

Gopnik, A., & Wellman, H. M. (1994). The theory theory. In S. A. Gelman & L. A. Hirschfeld (Eds.), *Mapping the mind: Domain specificity in cognition and culture; Based on a conference entitled "Cultural Knowledge and Domain Specificity," held in Ann Arbor, MI, Oct 13-16, 1990* (pp. 257-293). New York, NY, US: Cambridge University Press.

Griffiths, T.L., & Tenenbaum, J.B. (2001) Randomness and coincidences: Reconciling intuition and probability theory. *Proceedings of the 23rd Annual Conference of the Cognitive Science Society.*

Harris, P. L., German, T., & Mills, P. (1996). Children's use of counterfactual thinking in causal reasoning. *Cognition, 61*(3), 233-259.

Inagaki, K., & Hatano, G. (1993). Young children's understanding of the mind body distinction. *Child Development, 64*(3), 1534-1549.

Kalish, C. (1996). Causes and symptoms in preschoolers' conceptions of illness. *Child Development, 67*(4), 1647-1670.

Keil, F. C. (2003). Folkscience: Coarse interpretations of a complex reality. *Trends in Cognitive Sciences, 18*(5-6), 663-692.

Kuhn, D. (1989). Children and adults as intuitive scientists. *Psychological Review, 96*(4), 674-689.

Kuhn, D., Amsel, E. & O'Laughlin, M. (1988). *The development of scientific thinking skills.* Orlando, Florida: Academic.

Kushnir, T. (2003). *Seeing versus doing: The effect of direct intervention on preschooler's understanding of probabilistic causes.* Poster presented at the Biennial meeting of the Society for Research in Child Development, Tampa, FL.

Kushnir T. & Gopnik, A., (in press) Children infer causal strength from probabilities and interventions. *Psychological Science.*

Kushnir, T., Gopnik, A. & Schaefer, C. (2005, April). Children infer hidden causes from probabilistic evidence. Paper presented at the biennial meeting of the Society for Research in Child Development, Atlanta, GA.

Kushnir, T., Gopnik, A., Schulz, L., & Danks, D. (2003). *Inferring hidden causes.* Paper presented at the 25th Conference of the Cognitive Science Society.

Leslie, A. M., & Keeble, S. (1987). Do six-month-old infants perceive causality? *Cognition, 25*(3), 265-288.

Masnick , A. M. & Klahr, D. (2003). Error matters: An initial exploration of elementary school children's understanding of experimental error*. Journal of Cognition and Development 4*(1), 67-98.

Meltzoff, A. N. (1995). Understanding the intentions of others: Re-enactment of intended acts by 18-month-old children. *Developmental Psychology, 31*, 838-850.

Michotte, A. (1962). *The perception of causality.* New York: Basic Books. (Original work published 1946).

Novick, L. R., & Cheng, P. W. (2004). Assessing interactive causal influence. *Psychological Review, 111*(2), 455-485.
Pearl, J. (2000). *Causality*. New York: Oxford University Press.

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

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

Piaget, J. (1930). *The child's conception of physical causality*. London: Kegan Paul.

Rovee-Collier, C. (1980). Reactivation of infant memory. *Science 208*(4448), 1159-1161.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science, 274*(5294), 1926-1928.

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

Salmon, W. (1998). *Causality and explanation.* Oxford: Oxford University Press.

Schaefer, C., & Gopnik, A. (2003). *Causal reasoning in young children: The role of unobserved variables.* Paper presented at the Biennial meeting of the Society for Research in Child Development.

Schlottman, A. (2001). Perception versus knowledge of cause and effect in children: When seeing is believing. *Current Directions in Psychological Science 10*(4), 111-115.

Schulz, L. E., (2001, December). *Spinning wheels and bossy ones: Children, causal structure and the calculus of intervention*. Paper presented at the Causal Inference in Humans and Machines Workshop of the Neural Information Processing Systems annual meeting, Vancouver British Columbia.

Schulz, L. (2003). *The play's the thing: Interventions and causal inference.* Paper presented at the biennial meeting of the Society for Research in Child Development, Tampa, FL.

Schulz, L., & Gopnik, A. (2004). Causal learning across domains. *Developmental Psychology, 40*(2), 162-176.

Schulz, L., Gopnik, A., & Glymour, C. (in submission). Preschool children learn about causal structure from interventions.

Shanks, D. R. (1985). Forward and backward blocking in human contingency judgment. *Quarterly Journal of Experimental Psychology: Comparative & Physiological Psychology, 37*(1), 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, US: Academic Press, Inc.

Shultz, T. R. (1982). Rules of causal attribution. *Monographs of the Society for Research in Child Development, 47*(1), 1-51.

Sloman, S. A., & Lagnado, D. A. (2005). Do we "do"? *Cognitive Science, 29*(1), 5-39.

Sobel, D. M., Tenenbaum, J. B., & Gopnik, A. (2004). Children's causal inferences from indirect evidence: Backwards blocking and Bayesian reasoning in preschoolers. *Cognitive Science, 28,* 303-333.

Sommer, S., & Wehner, R. (2004). The ant's estimation of distance travelled: experiments with desert ants, Cataglyphis fortis. *Journal of comparative physiology A - neuroethology sensory neural behavioral physiology, 190*(1), 1-6.

Spelke, E. S., Breinlinger, K., Macomber, J., & Jacobson, K. (1992). Origins of knowledge. *Psychological Review, 99*(4), 605-632.

Spelke, E. S., Katz, G., Purcell, S. E., Ehrlich, S. M., & Breinlinger, K. (1994). Early knowledge of object motion: Continuity and inertia. *Cognition, 51*, 131-176.

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

Steyvers, M., Tenenbaum, J., Wagenmakers, E. J., & 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, and K. Obemayer (Eds.), *Advances in Neural Information Processing Systems 15*. Cambridge, MA: MIT Press.

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

Wasserman, E. A., Elek, S. M, Chatlosh, D. L., & Baker, A. G. (1993). Rating causal relations: Role of probability in judgments of response-outcome contingency. *Journal of Experimental Psychology: Learning, Memory and Cognition, 19*(1), 174-188.

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.; cognitive development in infancy; portions of this paper were initially reported at the biennial meeting of the society for research in child development, santa monica, CA, 1969.* (pp. 77-85). Hillsdale, NJ, England: Lawrence Erlbaum Associates, Inc.

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. New directions for child development, No. 75* (pp. 7-25). San Francisco, CA, US: Jossey-Bass/Pfeiffer.

Tenenbaum, J., & Griffiths, T. L. (2003). Theory-based causal inference. In S. Becker, S. Thrun & K. Obemayer (Eds.), *Advances in Neural Information Processing Systems 15* (pp. 67-74). Cambridge, MA: MIT Press.

Tenenbaum, J., & Niyogi, S. (2003). Learning causal laws. *Proceedings of the Twenty-Fifth Annual Conference of the Cognitive Science Society*.

Waldmann, M. R. (2001). Predictive versus diagnostic causal learning: Evidence from an overshadowing paradigm. *Psychonomic Bulletin & Review, 8,* 600-608.

Waldmann, M. R. (2000). Competition among causes but not effects in predictive and diagnostic learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26,* 53-76.

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

Wellman, H. M., Hickling, A. K., & Schult, C. A. (1997). Young children's psychological, physical, and biological explanations. *New directions for child development, 75*, 7-25.

Woodward, A. L. (1998). Infants selectively encode the goal object of an actor's reach. *Cognition, 69*, 1-34.

Woodward, A. L., Phillips, A. T., & Spelke, E. S. (1993). Infants' expectations about the motion of animate versus inanimate objects. Proceedings of the Fifteenth Annual Meeting of the Cognitive Science Society (pp. 1087-1091). Hillsdale, NJ: Erlbaum.

Woodward, J. (2003). *Making things happen: A theory of causal explanation*. New York: Oxford University Press.

Wu, M., & Cheng, P. W. (1999). Why causation need not follow from statistical association: Boundary conditions for the evaluation of generative and preventive causal powers. *Psychological Science, 10*(2), 92-97.