A. Gopnik, & C. Glymour (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*. Cambridge: Cambridge University Press. 117-132.

Causal maps and Bayes nets: a cognitive and computational account of theory-formation

Alison Gopnik and Clark Glymour

In this chapter, we outline a more precise cognitive and computational account of the 'theory theory' of cognitive development. Theories and theory-formation processes are cognitive systems that allow us to recover an accurate 'causal map' of the world: an abstract, coherent representation of the causal relations among events. This kind of knowledge can be perspicuously represented by the formalism of directed graphical causal models, or 'Bayes nets'. Human theory formation may involve similar computations.

1 The theory theory

Recently cognitive psychologists have argued that much of our adult knowledge, particularly our knowledge of the physical, biological and psychological world, consists of 'intuitive' or 'naïve' or 'folk' theories (Murphy and Medin, 1985; Rips, 1989). Similarly, cognitive developmentalists argue that children formulate and revise a succession of such intuitive theories (Carey, 1985; Gopnik, 1988; Keil, 1989; Wellman, 1990; Gopnik and Meltzoff, 1997; Wellman and Gelman, 1997). This idea, which we have called the 'theory theory', rests on an analogy between everyday knowledge and scientific theories. Advocates of the theory theory have drawn up lists of features that are shared by these two kinds of knowledge. These include static features of theories, such as their abstract, coherent, causal, counter-factual-supporting character; functional features of theories such as their ability to provide predictions, interpretations and explanations; and dynamic features such as theory changes in the light of new evidence (see Gopnik and Wellman, 1994; Gopnik and Meltzoff, 1997). The assumption behind

this work is that there are common cognitive structures and processes, common representations and rules, that underlie both everyday knowledge and scientific knowledge.

If this is true it should be possible to flesh out the nature of those cognitive structures and processes in more detail. Formulating the analogy between science and development has been an important first step, but, like all analogies, it is only a first step. Moreover, as with all analogies, there is a risk that we will end up in endless disputes about whether specific features of the two types of cognition are similar or different. Light waves aren't wet, after all, and natural selection takes place in an entirely different time-scale than artificial selection. Scientific theory change is different from child development in many ways – it is a social process, it involves a division of labour, and so on (see Faucher *et al.*, this volume). Rather than wrangling over whether these differences invalidate the analogy it would be more productive to describe in some detail the representations and rules that could underpin both these types of knowledge. Ideally, such an account should include ideas about the computational character of these representations and rules, and eventually even their neurological instantiation. This is the project that has been so successful in other areas of cognitive science, particularly vision science.

In this paper we will outline a more developed cognitive and computational account of the theory theory. In particular, we will argue that many everyday theories and everyday theory changes involve a type of representation we will call a 'causal map'. A causal map is an abstract representation of the causal relationships among kinds of objects and events in the world. Such relationships are not, for the most part, directly observable, but they can often be accurately inferred from observations. This includes both observations of patterns of contingency and correlation among events as well as observations of the effects of experimental interventions. We can think of everyday theories and theory-formation processes as cognitive systems that allow us to recover an accurate causal map of the world.

We will argue that this kind of knowledge can be perspicuously represented by the formalism of directed graphical causal models, more commonly known as Bayes nets (Pearl, 1988, 2000; Spirtes *et al.*, 1993). Interestingly, this formalism has its roots in work in the philosophy of science. It is the outcome, at least in part, of an attempt to

characterize the inductive inferences of science in a rigorous way. The formalism provides a natural way of representing causal relations, it allows for their use in prediction and experimental intervention, and, most significantly, it provides powerful tools for reliably inferring causal structures from patterns of evidence. In recent work in artificial intelligence, systems using Bayes nets can infer accurate, if often incomplete, accounts of causal structure from suitable correlational data. We will suggest that human causal inference and theory formation may involve more heuristic versions of similar computations.

2 Theory-formation and the causal inverse problem

The most successful theories in cognitive science have come from studies of perception. The visual system, whether human or robotic, has to recover and reconstruct three-dimensional information from the retinal (or fore-optic) image. One aspect of vision science is about how that reconstruction can be done computationally, and about how it is done in humans. Although accounts are very different in detail they share some general assumptions, in particular these: (1) Visual systems, whether human or automated, have an objective problem to solve: they need to discover how three-dimensional moving objects are located in space. (2) The data available are limited in particular ways. Organisms have no direct access to the external world. Rather, the external world causes a flow of information at the senses that is only indirectly related to the properties of the world itself. For example, the information at the retina is two-dimensional, while the world is three-dimensional. (3) Solutions must make implicit assumptions about the ways that objects in the world produce particular patterns – and successions of 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. But these assumptions are also contingent; if the assumptions are violated then the system will generate incorrect representations of the world (as in perceptual illusions). (See Palmer, 1999.)

We propose an analogous problem about discovering the causal structure of the environment. (1) There are causal facts, as objective as facts about objects, locations, and states of relative motion, used and evidenced in prediction, intervention and control, and partially revealed in correlations. (2) The data available are limited in

particular ways. Children and adults may observe associations they cannot control or manipulate; they may observe features they can only control or manipulate indirectly, through other objects or features; the associations they observe, with or without their own interventions, may involve an enormous number of features, only some of which are causally related. (3) Human beings have a theory-formation system, like the visual system, that recovers causal facts by making implicit assumptions about the causal structure of the environment. Those assumptions are contingent; where they are false, causal inference, whether in learning new causal relationships or in deploying old ones, may fail to get things right.

3 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 and rules to solve the spatial problem. In some cases, like the case of translating two-dimensional retinal information to three-dimensional representations, the kinds of representations and the rules that generate them may be relatively fixed and 'hard-wired'. However, other, more flexible, kinds of spatial representations are also used

In particular, since Tolman, cognitive scientists have suggested that organisms solve the spatial inverse problem by constructing 'cognitive maps' of the spatial environment (O'Keefe and Nadel, 1978; Gallistel, 1990). These cognitive maps provide animals with representations of the spatial relations among objects. Different species of animals, even closely related species, may use different types of cognitive map. There is some evidence suggesting the sorts of computations that animals use to construct spatial maps, and there is even some evidence about the neurological mechanisms that underpin those computations. In particular, O'Keefe and Nadel (1978) proposed that these mechanisms were located in the rat hippocampus.

There are several distinctive features of cognitive maps. First, such maps provide non-egocentric representations. Animals might navigate through space, and sometimes do, egocentrically, by keeping track of the changing spatial relations between their bodies and objects as they move through the spatial environment. In fact, however, 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 who 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 behaviourists, which depend on associations between external stimuli and the animal's own responses. This, of course, made Tolman one of the precursors of the cognitive revolution.

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 who knows the spatial layout of a maze can use that information to make many new inferences about objects in the maze. For example, the animal can conclude that if A is north of B, and B is north of C, then A will be north of C. The coherence of cognitive maps gives them their predictive power. An animal with a spatial map can make a much wider variety of predictions about where an object will be located, than can an animal restricted to egocentric spatial navigation. It also gives cognitive maps a kind of interpretive power; an animal with a spatial map can use the map to resolve ambiguous spatial information.

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. A rat moving towards a bait in a familiar maze is in a very different position than, say, a moth moving towards a lamp. The moth appears to be hard-wired to respond in set ways to particular stimuli in the environment. The rat in contrast moves in accordance with a learned representation of the maze. This also means that spatial cognitive maps may be defeasible. As an animal explores its environment and gains more information about it, it will alter and update its map of that environment. In fact, it is interesting that the hippocampus, which, in rats, seems to be particularly involved in spatial map-making, also seems particularly adapted for learning and memory. Of course, this general learning ability depends on innate learning mechanisms that are specialized and probably quite specific to the spatial domain such as, for example, 'dead reckoning' mechanisms (see Gallistel, 1990).

Our hypothesis is that human beings construct similar representations that capture the causal character of their environment. This capacity plays a crucial role in

the human solution to the causal inverse problem. These causal maps are what we refer to when we talk about everyday theories. Everyday theories are non-egocentric, abstract, coherent, learned representations of causal relations among events, and kinds of events, that allow causal predictions, interpretations and interventions.

Note that we are not proposing that we actually use spatial maps for the purpose of representing or acquiring causal knowledge, or that we somehow extend spatial representations through processes of metaphor or analogy. Rather we want to 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 with which it must in many cases interact. We also do not mean that knowledge of causal relations is developed entirely independently of knowledge of spatial facts, but that there are special problems about learning causal relationships, and special types of representations designed to solve those problems.

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, we 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 which 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. These sorts of specific hard-wired representations could capture particular important parts of the causal structure of the environment. This is precisely the proposal that Heider (1958) and Michotte (1962) made regarding the 'perception' of both physical and psychological causality. There is evidence for such representations even in young infants (Leslie and Keeble, 1987; Oakes and Cohen, 1990). Animals might also be hard-wired to detect specific kinds of causal relations that involve especially important events, such as the presence of food or pain. Such capacities appear to be involved in phenomena like classical conditioning or the Garcia effect, in which animals avoid food that leads to poisoning (Palmerino *et al.*, 1980).

Animals could also use a kind of egocentric causal navigation, they might

calculate the 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. Operant conditioning allows an animal to calculate the novel causal effects of its own actions on the world, and to take this information into account in future actions. More generally, trial-and-error learning seems to involve similar abilities for egocentric causal navigation.

Causal maps, however, would confer the same sort of advantages as spatial maps (Campbell, 1995). With a non-egocentric causal representation of the environment, an animal could predict the causal consequences of an action without actually having to perform it. The animal could 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. Similarly, an animal with a causal map could update the information in that map simply by observing causal interactions in the world, for example, by observing the causal consequences of another animal's actions, or by observing causal phenomena in the environment. The animal could then use that information to guide its own goal-directed actions. The coherence of causal maps allows a wide range of predictions. Just as an animal with a spatial map could make transitive spatial inferences (if A is north of B, and B is north of C, then A will be North of C) animals with causal maps could make transitive causal inferences (if A causes B, and B causes C, then A will cause C). A causal map would allow for a wide range of causal predictions and also allow a way of interpreting causally ambiguous data.

Since causal maps are learned they should give animals an opportunity to master new causal relations, not just whatever limited set might be 'hard-wired' perceptually. We would expect that animals would perpetually extend, change and update their causal maps just as they update their spatial maps.

It may well be that human beings are, in fact, the only animals that construct causal maps. In particular, there is striking recent evidence that chimpanzees, our closest primate relatives, rely much more heavily on egocentric trial and error causal learning. While chimpanzees are extremely adept at learning how to influence their environments, quite possibly more adept than human beings are, they have a great deal of difficulty appreciating causal relations among objects that are independent of their

own actions (Tomasello and Call, 1997; Povinelli, 2000). Chimpanzees are, at best, severely restricted in their ability to learn by observing the interventions of others, or by observing causal relations among objects and events.

4 Theories as causal maps

'Everyday' or 'folk' theories seem to have much of the character of causal maps. Such 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 actions. They postulate coherent relations among such objects and events which support a wide range of predictions, interpretations and interventions. Because of their causal character, they support counter-factual reasoning. Moreover, theories, like causal maps, are learned through our experience of and interaction with the world around us. Because of this, the theory theory has been especially prominent as a theory of cognitive development.

These are also features that unite everyday theories and scientific theories. While not all scientific theories are causal, causal claims and inferences do play a central role in most scientific theories (see Salmon, 1984; Cartwright, 1989). Scientific theories also involve learned, coherent, non-egocentric networks of causal claims which support prediction, interpretation, counter-factual reasoning and explanation. Moreover, when scientific theories are less concerned with causal structure it tends to be because these theories involve formal mathematical structure instead. However, this is also one way in which scientific theories appear to be unlike everyday theories. Scientific theories seem to be a-causal chiefly only in so far as they are formulated in explicit formal and mathematical terms. Everyday theories are rarely formulated in such terms.

Moreover, the idea of causal maps seems to capture the scope of 'theory theories' very well. The theory theory has been very successfully applied to our everyday knowledge of the physical, biological and psychological worlds. However, the theory theory does not seem to be as naturally applicable to other types of knowledge, for example, purely spatial knowledge, syntactic or phonological knowledge, musical knowledge, or mathematical knowledge. Nor does it apply to the much more loosely organized knowledge involved in empirical generalizations, scripts or associations (Gopnik and Meltzoff, 1997). But these types of knowledge also do not appear to

involve causal claims in the same way. Conversely some kinds of knowledge that do involve causal information, like the kinds of knowledge involved in operant or classical conditioning, do not seem to have the abstract, coherent, non-egocentric character of causal maps, and we would not want to say that this sort of knowledge was theoretical.

Some earlier accounts have proposed that theories, both scientific and everyday, should be cognitively represented as connectionist nets (Churchland, 1990) or as very generalized schemas (Giere, 1992). The difficulty with these proposals is that they seem to be too broad to capture what makes theories special; practically any kind of knowledge can be represented as nets or schemas. On the other hand, more modular accounts, such as that of Keil (1995) or Atran (this volume) have proposed that there are only a few specific explanatory schemas, roughly corresponding to the domains of physics, biology and psychology, that are used in everyday theories. These proposals do not seem to capture the wide range of explanations that develop in everyday life, nor the way that in cognitive development and in science we move back and forth among these domains. The idea of causal maps seems to capture both what is general and what is specific about everyday theories.

5 Bayes nets as causal maps

We propose, then, that children and adults construct causal maps: non-egocentric, abstract, coherent representations of causal relations among objects. An adequate representation of how such maps work must do three things: (1) It must show how they can be used to enable an agent to infer the presence of some features from other features of a system – it must allow for accurate predictions. (2) It must show how an agent is able to infer the consequences of her own or others actions – it must allow for appropriate interventions. And (3) it must show how causal knowledge can be learned from the agent's observations and actions.

There has recently been a great deal of computational work investigating such representations and mechanisms. The representations commonly called Bayes nets can model complex causal structures and generate appropriate predictions and interventions. Moreover, we can use Bayes nets to infer causal structure from patterns of associations, whether from passive observation or from action. 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 representation and inference algorithms allow one sometimes to uncover hidden unobserved causes, to disentangle complex interactions among causes, to make inferences about probabilistic causal relations and to generate counter-factuals (see Pearl, 1988, 2000; Spirtes *et al.*, 1993, 2000; Jordan, 1998).

This work has largely taken place in computer science, statistics and philosophy of science, and has typically been applied in one-shot 'data-mining' in a range of subjects, including space physics, mineralogy, economics, biology, epidemiology and chemistry. In these applications there is typically a large amount of data about many variables that might be related in a number of complex ways. Bayes net systems can automatically determine which underlying causal structures are compatible with the data, and which are ruled out. But these computational theories might also provide important suggestions about how human beings, and particularly young children, recover and represent causal information. Work in computer vision has provided important clues about the nature of the visual system, and this work might provide similar clues about the nature of the theory formation system. Causal maps might be a kind of Bayes net.

6 'Screening off' in causal inference

Bayes net systems are elaborations of a much simpler kind of causal inference, long discussed in the philosophy of science literature. Causal relations in the world lead to certain characteristic patterns of events. In particular, if A causes B, then the occurrence of A will change the probability that B will occur. We might think that this could provide us with a way of solving the causal inverse problem. When we see that A is correlated with B – that is, that the probability if A is consistently related to the probability of B – we can conclude that A caused B (or vice-versa).

But there is a problem. The problem is that other events might also be causally related to B. For example, some other event C might be a common cause of both A and B. A doesn't cause B but whenever C occurs both A and B will occur together. Clearly, what we need in these cases is to have some way of considering the probability of A and B relative to the probability of C. Many years ago the philosopher of science Hans Reichenbach proposed one natural way of doing this, which he called 'screening

off' (Reichenbach, 1956). We can represent this sort of reasoning formally as follows. If A, B and C are the only variables and A is only correlated with B conditional on C, C 'screens off' A as a cause of B – C rather than A is the cause. If A is correlated with B independent of C, then C does not screen off A and A causes B. This sort of 'screening off' reasoning is ubiquitous in science. In experimental design we control for events that we think might be confounding causes. In observational studies we use techniques like partial correlation to control for confounding causes. The trouble with the reasoning we've described so far is that it's limited to these rather simple cases. But, of course, in real life events may involve causal interactions among dozens or even hundreds of variables rather than just three. And the relations among variables may be much more complicated than either of the simple structures we described above. The causal relations among variables may also have a variety of different structures – A might be linearly related to B, or there might be other more complicated functions relating A and B, or A might inhibit B rather than facilitating it. The causal relations might involve Boolean combinations of causes. Or A and B together might cause C, though either event by itself would be insufficient. And there might be other unobserved hidden variables, ones we don't know about, that are responsible for patterns of correlation. Moreover, in the real world, even when the causal relations we are uncovering are deterministic, the evidence for them is almost invariably noisy and probabilistic. Is there a way to generalize the 'screening off' reasoning we use in the simple cases to these more complicated ones? Would a similar method explain how more complicated causal relations might be learned? The Bayes net formalism provides such a method.

7 Bayes nets and their uses

Bayes nets are directed graphs, like the one shown below. The nodes or vertices of the graph represent variables, whose values are features or properties of the system, or collections of systems to which the net applies. 'Colour,' for example, might be a variable with many possible values; 'weight' might be a variable with two values, heavy and light, or with a continuum of values. When Bayes nets are given a causal interpretation, a directed edge from one node or variable to another – X to Y, for example – says that an intervention that varies the value of X but otherwise does not

alter the causal relations among the variables will change the value of Y. In short, changing X will cause Y to change.

For each value of each variable, a probability is assigned, subject to a fundamental rule, the Markov assumption. The Markov assumption is a generalization of the 'screening off' property we just described. It says that if the edges of the graphs represent causal relations, then there will only be some patterns of probabilities of the variables, and not others. The Markov assumption constrains the probabilities that can be associated with a network. It says that the various possible values of any variable, X, are independent of the values of any set of variables in the network that does not contain an effect (a descendant of X), conditional on the values of the parents of X. So, for example, applied to the directed graph 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}.

Figure 1 about here

Bayes nets allow causal predictions. Information that a system has some property or properties often changes the probabilities of other features of the system. The information that something moves spontaneously, for example, may change the probability that it is animate. Such changes are represented in Bayes nets by the conditional probability of values of a variable given values for another variable or variables. Bayes net representations simplify such calculations in many cases. In the network above, the probability of a value of X conditional on a value of R may be calculated from the values of p(R | S), p(S) and p(X | S). (See Pearl, 1988.) This allows us to predict the value of X if we know the value of R. Bayes-nets also provide a natural way of assessing and representing counter-factual claims (Pearl, 2000).

In planning we specifically predict the outcome of an action. The probabilities for various outcomes of an action that directly alters a feature are not necessarily the same as the probabilities of those outcomes conditional on that altered feature. Suppose R in the graph above has two values, say red and pink. Because the value of S influences R, the conditional probabilities of values of S given that R = red will be different from the conditional probabilities of values of S given that R = pink. Because S influences X, the probabilities of values of X will also be different on the two values

of R. *Observing* the value of R gives information about the value of X. But R has no influence on S or X, either direct or indirect, so if the causal relations are as depicted, acting from outside the causal relations represented in the diagram to change the value of R will do nothing to change the value of S or X. It is possible to compute over any Bayes network which variables will be indirectly altered by an action or intervention that directly changes the value of another variable. It is also possible to compute the probabilities that indirectly result from the intervention. These calculations are sometimes possible even when the Bayes net is an incomplete representation of the causal relations. (See Pearl and Verma, 1991; Spirtes *et al.*, 1993; Glymour and Cooper, 1999.)

Bayes nets thus have two of the features that are needed for applying causal maps: they permit prediction from observations, and they permit prediction of the effects of actions. With an accurate causal map – that is, the correct Bayes net representation – we can accurately predict that y will happen when x happens, or that a particular change in x will lead to a particular change in y, even when the causal relations we are considering are quite complex. Similarly, we can accurately predict that if we intervene to change x then we will being about a change in y. Bayes nets have another feature critical to cognition: they can be learned.

8 Learning Bayes nets

In 'data mining' applications Bayes nets have to be inferred from uncontrolled observations of variables. To do this, the Markov assumption is usually supplemented by further assumptions. The additional assumptions required depend on the learning procedure. (A detailed survey of several learning algorithms is given in essays in Glymour and Cooper, 1999.) One family of algorithms uses Bayes Theorem to learn Bayes nets. Another class of algorithms learns the graphical structure of Bayes nets entirely from independence and conditional independence relations among variables in the data, and requires a single additional assumption. We will describe some features of the latter family of algorithms.

The additional assumption required is *faithfulness*: the independence and conditional independence relations among the variables whose causal relations are described by a Bayes net must all be consequences of the Markov assumption applied

to that network. For example, in the figure above, it is possible to assign probabilities so that S and X are independent, although the Markov assumption implies no such independence. (We can arrange the probabilities so that the association of S and X due to the influence of S on X is exactly cancelled by the association of S and X due to the influence of Y on both of them.) The faithfulness assumption rules out such probability arrangements.

The faithfulness assumption is essentially a simplicity assumption. It is at least logically possible that the contingencies among various causes could be randomly arranged in a way that would 'fool' a system that used the causal Markov condition. The faithfulness condition assumes that in the real world such sinister coincidences will not take place.

The learning algorithms for Bayes nets are designed to be used either with or without background knowledge of specific kinds. In addition to the Markov assumption and the faithfulness assumption we may add other assumptions about how causes are related to events. For example, an agent may, and a child typically will, know the time order in which events occurred, and may believe that some causal relations are impossible and others certain. Information of that sort is used by the algorithms. For example, suppose the child, or whatever agent, knows that events of kind A come before events of kind B which come before events of kind C. Suppose the true structure were as represented in Figure 2. Given data in which A, B and C are all associated, a typical Bayes net learning algorithm such as the TETRAD II 'Build' procedure (Scheines *et al.*, 1994) would use the information that A precedes B and C to test only whether B and C are independent conditional on A. Finding that conditional independence, the algorithm will conjecture the structure in Figure 2. No other structure is consistent with the associations, the conditional independence, the time order, and the Markov and faithfulness assumptions.

Figure 2 about here

Given the Markov and faithfulness assumptions, then, we can construct algorithms that will arrive at the correct causal structure if they are given information about the contingencies among events. These systems can learn about causal structure

from observations and interventions.

9 Bayes nets and adults

Human adults seem to have causal maps that go beyond the causal representations of classical or operant conditioning. Is there any empirical evidence that these maps also involve Bayes net-like representations? In fact, there is some evidence that adults make causal judgments in a way that respects the assumptions of a Bayes net formalism. There is a long literature, going back to Kelley in social psychology, about the way that adults perform a kind of causal 'discounting' (Kelley, 1973). Adults seem to unconsciously consider the relationships among possible causes, that is, to consider alternative causal graphs, when they decide how much causal influence one event has on another event. In particular, Patricia Cheng's recent 'causal power' theory turns out to be equivalent to a particular common parameterisation of causal graphs in Bayes net theories (Cheng, 1997). Cheng's theory, which was empirically motivated and developed independently of the Bayes net work, makes the same assumptions about the relation between causal graphs and probabilities that are made in these AI models (Glymour, in press). In effect, Cheng's work suggests that adults may use a Bayes net-like representation to make causal predictions.

While Bayes nets provide tools for prediction and intervention, they also admit algorithms for learning new causal relations from patterns of correlation. Interestingly, however, there is little work on how adults learn about new causal relations. This is probably because adults rely overwhelmingly on their prior causal knowledge of the world in making causal judgments. They already have rich, powerful, well-confirmed theoretical assumptions about what will cause what. Because of the enormous causal knowledge adults bring with them, experimentation on adult causal learning is virtually forced to imaginary or artificial scenarios to separate learning from the application of background knowledge. In everyday life, adults may rarely be motivated to revise their earlier causal knowledge or construct new knowledge of a general kind (of course adults learn new causal particulars every day). The cognitive problem for adults is to apply that knowledge appropriately in particular situations.

The situation is very different for children. Interestingly it is also different in the special conditions in which adult human beings do science. By definition scientific

inquiry is precisely about revising old causal knowledge and constructing new causal knowledge – science is quintessentially about learning. It is no coincidence that work on causal inference and in particular the Bayes net formalism has largely been done by philosophers of science, rather than cognitive psychologists. Human capacities for learning new causal facts about the world may be marginal for understanding much everyday adult cognition, but they are central for understanding scientific cognition.

10 Bayes nets and learning

We propose that the best place to look for powerful and generalized causal learning mechanisms, learning of the sort that might be supported by Bayes net algorithms, is in human children. Unlike adults, children cannot just rely on prior knowledge about causal relations. Prior knowledge isn't prior until after you've acquired it. And empirically, we have evidence that massive amounts of learning, particularly causal learning, take place in childhood. Indeed, in some respects the cognitive agenda for children is the reverse of the agenda for adults. Children are largely protected from the exigencies of acting swiftly and efficiently on prior knowledge, adults take those actions for them. But children do have to learn a remarkable amount of new information, in a relatively short time, with limited but abundant evidence.

Moreover, unlike non-human animals, children's learning must extend well beyond the limited set of causal relations that involve adaptively important mechanisms or involve the effects of one's own actions. Both human adults and children themselves have a large store of information about causal relations that do not involve positive or negative reinforcement and are not the result of the actions (this, of course, was one of the lessons of the cognitive revolution).

We have prima facie evidence that children do, in fact, learn an almost incredible amount about the causal structure of the world around them. That is the evidence that supports the theory theory in general. Similarly, we have prima facie evidence that, in the special circumstances of science, adult human minds can uncover new causal structure in the world. In the case of science, philosophers for many years simply abandoned the hope of uncovering a logic for this sort of causal induction. In cognitive psychology, the computational models of learning that have been proposed have been either the highly constrained 'parameter setting' models of modularity

theories (see, e.g., Pinker, 1984), or the highly unconstrained and domain-general regularity detection of connectionist modelling (see, e.g., Elman *et al.*, 1996). Neither of these alternatives has been satisfactory as a way of explaining children's learning of everyday theories, or scientific change. Bayes net representations and computations provide a promising alternative. Such representations might play an important role in the acquisition of coherent causal knowledge.

In several recent studies, we have begun to show that children as young as two years old, in fact, do swiftly and accurately learn new causal relations – they create new causal maps. They do so even when they have not themselves intervened to bring about an effect, and when they could not have known about the relation through an evolutionarily determined module or through prior knowledge. We present children with a machine, 'the blicket detector', that lights up and plays music when some objects but not others are placed upon it. Children observe the contingencies between the objects and the effects and have to infer which objects are 'blickets'. That is, they have to discover which objects have this new causal power. Children as young as two years old swiftly and accurately make these inferences. They identify which objects are blickets and understand their causal powers (Gopnik and Sobel, 2000; Nazzi and Gopnik, 2000).

Moreover, they do so in ways that respect the assumptions of the Bayes Net formalism. In particular, even very young children use a form of 'screening-off' reasoning to solve these problems (Gopnik *et al.*, 2000).

There are some important caveats here. Whenever we apply computational work to psychological phenomena we have no guarantee that the human mind will behave in the same way as a computer. We even have important reasons to think that the two will be different. Clearly, the computations we propose would be performed unconsciously both by children and adults. (Since the three-year-olds we are considering are still unable to consciously add two plus two it is rather unlikely that they would be consciously computing exact conditional probabilities). Moreover, it is likely – indeed, almost certain – that human children rely more heavily on prior knowledge and on various heuristics than the current Bayes net 'data mining' systems do.

Nevertheless we would again draw an analogy to our understanding of vision

and spatial cognition. In this area there has been a thoroughgoing and extremely productive two-way interaction between computational and psychological work. While computational vision systems clearly have different strengths and weaknesses than human vision, they have proved to be surprisingly informative. Moreover, and rather surprisingly, the human visual system often turns out to use close to optimal procedures for solving the spatial inverse problem.

The program we propose is therefore not to theorize that children or scientists are optimal data miners, but rather to investigate in general how human minds learn causal maps, and how much (and, possibly, how little) their learning processes accord with Bayes net assumptions and heuristics.

11 Conclusion.

In a book published a mere three years ago the first author of this paper expressed pessimistic sentiments about the prospect of a computational account of everyday theory-formation and change. 'Far too often in the past psychologists have been willing to abandon their own autonomous theorizing because of some infatuation with the current account of computation and neurology. We wake up one morning and discover that the account that looked so promising and scientific – S-R connections, gestaltist field theory, Hebbian cell assemblies – has vanished and we have spent another couple of decades trying to accommodate our psychological theories to it. We should summon up our self-esteem and be more stand-offish in future' (Gopnik and Meltzoff, 1997). We would not entirely eschew that advice. Pessimism may, of course, still turn out to be justified – what we have presented here is a hypothesis and a research program rather than a detailed and well-confirmed theory. Moreover, we would emphasize that, as in the case of computer vision, we think the computational accounts have as much to learn from the psychological findings as vice-versa. Nevertheless, sometimes it is worth living dangerously. We hope that this set of ideas will eventually lead, not to another infatuation, but to a mutually rewarding relationship between cognition and computation.

Acknowledgements

Versions of this paper were presented at the International Congress on Logic, Methodology and Philosophy of Science, Cracow, Poland, in August, 1999; and at seminars at the University of Chicago, the California Institute of Technology, and the Cognitive Science program and Department of Statistics at Berkeley. We are grateful to those who commented. Conversations with Steve Palmer, Lucy Jacobs, and Andrew Meltzoff played a major role in shaping these ideas. John Campbell and Peter Godfrey-Smith also read drafts of the paper and made very helpful suggestions.

Figure 1. A causal graph

Figure 2. The true structure of a causal graph