A. Gopnik & L. Schulz (2007). Introduction. To A. Gopnik & L. Schulz (eds.). *Causal learning: Psychology, philosophy, and computation*. New York: Oxford University Press, 2007, 358 pp.

Causal Bayes Nets for Dummies, The Psychology of Causal Inference for Nerds — A Correspondence

Alison Gopnik

Dept. of Psychology

University of California at Berkeley

Laura Schulz

Dept. of Brain and Cognitive Sciences

M.I.T.

From: mherskovits@psych.ucarcadia.arcadia.edu

To: brook_russell@turing.carnegietech.edu

Hi Brook,

We haven't met but I'm writing about this series of workshops on Causal Learning that my advisor and yours have cooked up for this year at the Center in Stanford. My advisor has gone completely meshuggena over this causal Bayes nets stuff and is

insisting that I go to this conference (on the pittance that supports graduate researchers) and that I learn everything there is to know about the philosophy and computation of causal learning. But every time I look at one of the papers all I see are unintelligible sentences like this: For any variable R in the directed graph, the graph represents the proposition that for any set S of variables in the graph, (not containing any descendants of R) R is jointly independent of the variables in S conditional on any set of values of the variables that are parents of R!

Let me give you a brief sense of where I'm coming from, as we say in mellow Arcadia (though I'm a New Yorker myself). I went to Public School 164 and did my undergraduate degree in cognitive science at the City University of Brooklyn and I've always thought that the problem of how we learn about the world was the most central and interesting question cognitive science could ask. That's why I became a developmental psychologist. But I'm suspicious about whether philosophy and computation have much to offer. The history of cognitive development, and the study of learning more generally, has been a history of theoretical answers that didn't really fit the phenomena, and empirical phenomena that didn't really fit the theories. What we empirical psychologists see is that learners infer abstract, structured hierarchical representations of the world. And those representations are true -- they really do get us to a better picture of the world. But the data that actually reach us from

the world are incomplete, fragmented, probabilistic and concrete. So the baffling thing for psychologists has been how we could get from that kind of data to those kinds of representations.

But the philosophers and computationalists keep telling us that the kind of learning we developmentalists see every day is nothing but an illusion! The Platonic (read Cartesian, read Chomskyan, read Spelkean) view has been that although we seem to infer structure from data, actually the structure was there all along. Insofar as our representations are accurate, it is because of a long phylogenetic evolutionary history, not a brief ontogenetic inferential one. And there is no real learning involved in development but only triggering or enrichment.

The Aristotelian (read Lockean, read behaviorist, read connectionist) view has been that although it looks as if we are building abstract veridical representations, really all we are doing is summarizing and associating bits of data. Accuracy is beside the point, associationistic processes just let us muddle through with the right responses to the right stimuli. There aren't really any abstract representations, just distributed collections of particular input-output links.

So all the philosophers and computationalists seem to be doing, on either side, is to tell us empirical developmental psychologists not to believe our eyes. Actually, I think Gopnik puts it quite well in her book about theory-formation (she

does tend to let her conclusions outstrip her data, but she sure has an ear for a slogan). "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, Gestaltian field theory, Hebbian cell-assemblies, has vanished and we have spent another couple of decades trying to accommodate our psychological theories to it. We think we should summon up our self-esteem and be more stand-offish in the future. Any implementations of psychological theories, either computational or neurological, will first depend on properly psychological accounts of psychological phenomena."

But anyway, although I've argued and argued my advisor is still insisting that I go to this thing. And it sounds like you're in the same boat. So I'm writing to you with a deal – How about a tutorial swap? You show me yours and I'll show you mine ;) – I mean, I'll tell you all about causal learning in psychology if you'll explain those goddamned Directed Acyclic Graphs in plain English words? So how 'bout it?

All best, Morgan Herskovits

From: brook_russell@turing.carnegietech.edu

To: mherskovits@psych.ucarcadia.arcadia.edu

My dear Morgan,

Thank you for your letter of the 21st. I can't say that we seem to have much else in common, but apparently your advisor matches mine in dotty obstinacy. He is insisting that I read all this barbaric and incomprehensible stuff about Subjects and Methods. And worse, it appears that quite a few of the Subjects appear to be between 30.1 months and 40.8 months—sprogs in short! But what on earth Methods for Sprogs are supposed to have to do with discovering normatively reliable methods for causal inference I can't imagine. And he is also insisting that I attend these workshops.

I can't say I caught all your references. Plato certainly but Spelke? Gopnik? (and what ghastly names). However, I completely agree with you about the lack of connexion between our two enterprises. The philosopher of science Clark Glymour put it very well, I think, in his critique of cognitive theories of science appropriately called "Invasion of the Mind Snatchers" – the idea that theories are something you would find in somebody's head, rather than being abstract mathematical objects, is an idea fit only for Ichabod Crane.

My own work began in my undergraduate days at Oxford, as an attempt at a conceptual analysis of causation. (I also am a public school product by the way –

though I find the idea of numbered public schools rather puzzling – would Eton or Harrow get a lower number on your American scheme?) The conceptual in philosophy, of course, is only a faux amie of the conceptual in psychology. In philosophy we want to know what causation IS in all conceivable circumstances, not what a few mere mortals (let alone sprogs!) think that it is. There is a long history in philosophy of trying to develop an analytic definition of causation through the method of examples and counter-examples – philosophers give examples of cases in which everyone agrees that X causes Y and then try to find some generalization that will capture those examples. And then other philosophers find examples that fit the definitions but don't seem to be causal or vice-versa.

I was working on counterexamples of quadruple countervailing causal prevention (you know the sort of thing where one assassin tries to stop another assassin but first poison is slipped in the antidote and then a brick hits a wooden board before the king can brake for the stop sign), But I was beginning to find it all rather discouraging when finally, my maths tutor put me on to the theory of causal graphical models, and it came to me as a revelation.

You see, causal graphical models are to causation as geometry is to space. Rather than providing a reductive definition of causation they instead provide a formal mathematical framework that captures important regularities in causal facts,

just as the mathematical structure of geometry captures important spatial regularities. Causal graphical models capture just the right kind of asymmetries in causal relations, allow one to generate the appropriate predictions about conditional probabilities and interventions, and perhaps most significantly of all discriminate between conditional probabilities and interventions and counterfactuals. So I decided to move to Carnegie Tech for graduate school and work on some of the many unsolved problems the formalism poses.

Imagine my shock, then when my advisor, a philosopher of science notorious for the austerity and rigor of his views on just about everything, began insisting that I read psychology and, worse, child psychology! Because, of course, it is obvious that even sophisticated adults are unable to handle even the simplest problems involving causality or probability. The undergraduate students at Carnegie Tech, for example, who, while admittedly handicapped by an American secondary school education, are among the brightest and best, are quite hopeless at these computations. Anyone who has, for their sins, had to teach introductory statistics is aware of that. So how could mere sprogs of three or four years be expected to use anything like Bayes net learning algorithms? They are I understand, inept at even quite elementary differential integration problems, and have, at best, only the most primitive understanding of basic linear algebra.

However, one of the benefits of an Oxford education is the training it provides in possessing a deep and thorough knowledge of the most recondite subjects based on a brief weekly perusal of the Times Literary Supplement. So I will, in fact, be grateful for a (preferably equally brief) summary of this work. And in return I will do my best to give you an extremely simple introduction to Causal Bayes Nets (see Attached))Yours very truly,

Brook Russell

Attachment 1: Causal Bayes Nets for Dummies

Causal Bayes Nets. Causal directed graphical models, or causal Bayes nets, have been developed in the philosophy of science and statistical literature over the last fifteen years (Glymour 2001; Pearl 1988, 2000; Spirtes et al. 1993.) Scientists seem to infer theories about the causal structure of the world from patterns of evidence. But philosophers of science found it very difficult to explain how these inferences are possible. Although classical logic could provide a formal account of deductive inferences, it was much more difficult to provide an inductive logic – an account of how evidence could confirm theories. One reason is that deductive logic deals in certainties but inductive inference is always a matter of probabilities – acquiring

more evidence for a hypothesis makes the hypothesis more likely, but there is always the possibility that it will be overturned. An even more difficult question was what philosophers of science called "the logic of discovery". Again the conventional wisdom, going back to Karl Popper, was that particular hypotheses could be proposed and could be falsified (definitely) or confirmed (tentatively). But the origins of those hypotheses were mysterious – there was no way of explaining how the evidence itself could generate a hypothesis.

Causal Bayes nets provide a kind of logic of inductive inference and discovery. They do so, at least, for one type of inference that is particularly important in scientific theory-formation. Many scientific hypotheses involve the causal structure of the world. Scientists infer causal structure by observing the patterns of conditional probability among events (as in statistical analysis), by examining the consequences of interventions (as in experiments) or, usually, by combining the two types of evidence. Causal Bayes nets formalize these kinds of inferences.

In causal Bayes nets, causal hypotheses are represented by directed acyclic graphs like the one below. The graphs consist of variables, representing types of events or states of the world, and directed edges (arrows) representing the direct causal relations between those variables (see figure 1). The variables can be discrete (like school grade) or continuous (like weight), they can be binary (like "having eyes"

or "not having eyes") or take a range of values (like color). Similarly, the direct causal relations can have many forms; they can be deterministic or probabilistic, generative or inhibitory, linear or non-linear. The exact specification of the nature of these relations is called the "parameterization" of the graph. In most applications of the formalism we assume that the graphs are acyclic – an arrow can't feed back on itself. However, there are some generalizations of the formalism to cyclic cases.

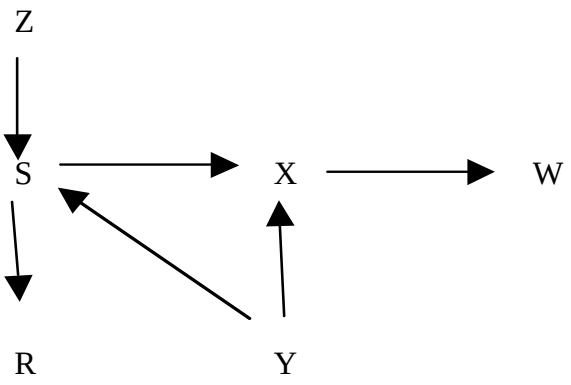

Figure 1: A causal Bayes net

Causal structure and conditional probabilities. The Bayes net formalism makes systematic connections between the causal hypotheses that are represented by the graphs and particular patterns of evidence. The structure of a causal graph con-

strains the conditional probabilities among the variables in that graph, no matter what the variables are or what the parameterization of the graph is. In particular, it constrains the conditional independencies among those variables. Given a particular causal structure, only some patterns of conditional independence will occur among the variables.

Conditional and unconditional dependence and independence can be defined mathematically. Two discrete variables X and Y are unconditionally independent in probability if and only if for every value x of X and y of Y the probability of x and y occurring together equals the unconditional probability of x multiplied by the unconditional probability of y. That is p (x & y) = p (x) * p (y.).Two variables are independent in probability conditional on some third variable Z if and only if p (x, y | z) = p (x | z) * p (y | z). That is for every value x,y, and z of X, Y and Z the probability of x and y given z equals the probability of x given z multiplied by the probability of y given z. This definition can be extended to continuous variables. When we say three variables x, y and z are correlated we mean that they are dependent in probability. When we say they x and y are correlated but that that correlation disappears when z is partialled out we mean that x and y are independent in probability conditional on z.

The structure of the causal graph puts constraints on these patterns of probability among the variables. These constraints can be captured by a single formal assumption, the Causal Markov Assumption as follows:

The Causal Markov Assumption: For any variable X in an acyclic causal graph, X is independent of all other variables in the graph (except for its own direct and indirect effects) conditional on its own direct causes.

If we make further assumptions about the parameterization of the graph, that is about the particular nature of the causal relations among the variables, we can constrain the kinds of inferences we make still further. For example, if we assume that each cause independently has a certain power to bring about an effect, and that that power leads to a certain likelihood of the effect given the cause, we can further constrain the patterns of conditional probability among causes and effects. This is a common assumption in studies of human causal learning. The Causal Markov assumption, however, applies to all parameterizations.

To illustrate this consider a simple causal problem that is far too common for academics who attend many learned conferences. Suppose that I notice that I often can't sleep when I've been to a party and drunk lots of wine. Partying (P) and insomnia (I) covary, and so do wine (W) and insomnia (I). There are at least two pos-

sibilities about the relations among these variables which I can represent by two simple causal graphs 1) a chain P → W→ I, or 2) a common cause structure I ← P → W. Maybe parties lead me to drink and wine keeps me up; maybe parties both keep me up and lead me to drink. The covariation among the variables by itself is consistent with both these structures.

You can discriminate between these two graphs by looking at the patterns of conditional probability among the three variables. Suppose you keep track of all the times you drink and party and examine the effects on your insomnia. If graph 1 is correct, then you should observe that you are more likely to have insomnia when you drink wine, whether or not you party. If, instead graph 2 is correct you will observe that, regardless of how much or how little wine you drink, you are only more likely to have insomnia when you go to a party.

More formally, if graph #1 is right, and there is a causal chain that goes from parties to wine to insomnia, then $I \perp P \mid W$ – the probability of insomnia occurring is independent (in probability) of the probability of party-going occurring conditional on the occurrence of wine-drinking If graph #2 is right, and parties are a common cause of wine and insomnia, then $I \perp W \mid P$ – the probability of wine-drinking occurring is independent (in probability) of the probability of insomnia occurring conditional on the occurrence of party-going.

The philosopher of science Hans Reichenbach long ago pointed out these consistent relations between conditional independence and causal structure and talked about them in terms of "screening-off". When there is a chain going from partying to wine to insomnia, the wine "screens off" insomnia from the influence of partying, when partying directly causes both wine and insomnia, wine does not screen-off insomnia from partying – partying leads to insomnia directly. But partying does "screen off" insomnia from the effects of wine. The causal Markov assumption generalizes this "screening-off" principle to all acyclic causal graphs.

Thus if we know the structure of the graph, and know the values of some of the variables in the graph, we can make consistent predictions about the conditional probability of other variables. In fact, the first applications of Bayes nets involved predicting conditional probabilities (Pearl 1988). Many real life inferences involve complex combinations of conditional probabilities among variables – consider a medical expert, for example, trying to predict one set of symptoms from another set. Trying to predict all the combinations of conditional probabilities rapidly becomes an exponentially complicated problem. Computer scientists were trying to find a tractable way to calculate these conditional probabilities, and discovered that representing the variables in a directed graph allowed them to do this. The graph allowed computer scientists to "read off" quite complicated patterns of conditional depend-

ence among variables. The first applications of Bayes nets treated the graphs as calculation devices -- summaries of the conditional probabilities among events.

Bayes nets and interventions. Why think of these graphs as representations of causal relations among variables, rather than simply thinking of them as a convenient way to represent the probabilities of variables? The earlier Bayes net iterations were confined to techniques for predicting some probabilities from others. However, the development of causal Bayes net algorithms also allows us to determine what will happen when we intervene from outside to change the value of a particular variable. When two variables are genuinely related in a causal way then, holding other variables constant, intervening to change one variable should change the other. Indeed, philosophers have recently argued that this is just what it means for two variables to be causally related (Woodward, 2003).

Predictions about probabilities may be quite different from predictions about interventions. For example, in a common cause structure like 2 above, we will indeed be able to predict something about the value of insomnia from the value of wine. If that structure is the correct one, knowing that someone drank wine will indeed make you more likely to predict that they will have insomnia (since drinking wine is correlated with partying, which leads to insomnia). But intervening on their

wine-drinking, forbidding them from drinking, for example, will have no effect on their insomnia. Only intervening on partying will do that.

The Bayes net formalism captures these relations between causation, intervention and conditional probability through a second assumption, an assumption about how interventions should be represented in the graph.

The Intervention Assumption: A variable I is an intervention on a variable X in a causal graph if and only if 1) I is exogenous (that is, it is not caused by any other variables in the graph) 2) directly fixes the value of X to x and 3) does not affect the values of any other variables in the graph except through its influence on X.

Given this assumption we can accurately predict the effects of interventions on particular variables in a graph on other variables. (We can also sometimes make accurate predictions about the effects of interventions that don't meet all these conditions). In causal Bayes nets, interventions systematically alter the nature of the graph they intervene on, and these systematic alterations follow directly from the formalism itself. In particular, when an external intervention fixes the value of a variable it also eliminates the causal influence of other variables on that variable. If I simply decide to stop drinking wine, my intervention alone will determine the value

of wine-drinking; partying will no longer have any effect. This can be represented by replacing the original graph with an altered graph in which arrows directed into the intervened upon variable are eliminated (Judea Pearl vividly refers to this process as graph surgery (Pearl 2000)). The conditional dependencies among the variables after the intervention can be read off from this altered graph.

Suppose, for example, I want to know what I can do to prevent my insomnia. Should I sit in my room alone, but continue to drink when I want to or go to parties just the same but stick to Perrier? I can calculate the effects of such interventions on each of the causal structures, using "graph surgery" and predict the results. I will get different results from these interventions depending on the true causal structure (solitary drinking will lead to insomnia, and sober partying won't for graph 1, sober partying will lead to insomnia and solitary drinking won't for graph 2

Exactly the same inferential apparatus can be used to generate counterfactual predictions. Suppose I want to ask what would have happened, had things been otherwise. If I had refrained from wine at all those conferences would my life, or at least my insomnia, have been better? Graph surgery will answer this question too. Just as in an intervention a counterfactual "fixes" the value of certain variables and allows you to infer the consequences.

A central aspect of causal Bayes nets, indeed the thing that makes them causal, is that they allow us to freely go back and forth from evidence about observed probabilities to inferences about interventions and vice-versa.

These two assumptions, then, allow us to take a particular causal structure and accurately predict the conditional probabilities of events, and also the consequences of interventions on those events, from that structure.

Bayes nets and learning

We just saw that knowing the causal structure let's us make the right predictions about interventions and probabilities. We can also use this fact to learn causal structure from the evidence of interventions and probabilities.

Lets go back to the wine-insomnia example. You could distinguish between these graphs either by intervention or observation. You could for instance, hold partying constant (always partying or never partying) and vary whether or not you drunk wine; or you could hold drinking constant (always drinking or never drinking) and vary whether or not you partied. In either case, you could observe the effect on your sleep. If drinking affects your sleep when partying is held constant, but partying has no effect on your sleep when drinking is held constant, you could conclude that graph 1 is correct. Such reasoning underlies the logic of experimental design in science.

You could also, however, simply observe the relative frequencies of the three events. If you notice that you are more likely to have insomnia when you drink wine, whether or not you party, you can infer that graph 1 is correct. If you observe that, regardless of how much or how little wine you drink, you are only more likely to have insomnia when you go to a party, you will opt instead for graph 2. These inferences reflect the logic of correlational statistics in science. In effect, what you did was to "partial out" the effects of partying on the wine/insomnia correlation, and draw a causal conclusion as a result.

This type of learning, however, requires an additional assumption, The assumption is that the patterns of dependence and independence we see among the variables really are the result of the causal relations among them. Suppose that wine actually makes you sleepy instead of keeping you awake. But it just happens to be the case that this influence of wine on insomnia is perfectly canceled out by the countervailing exciting influence of parties. We will incorrectly conclude that there are no causal relations between the three variables. We need to assume that these sinister coincidences will not occur. Formally, this is called the Faithfulness assumption.

The Faithfulness Assumption: In the joint distribution on the variables in the graph, all conditional independencies are consequences of the Markov as-

sumption applied to the graph.

Given the Faithfulness assumption, it is possible to infer complex causal structure from patterns of conditional probability and intervention (Glymour & Cooper, 1999; Spirtes et al., 1993). Computationally tractable learning algorithms have been designed to accomplish this task and have been extensively applied in a range of disciplines (e.g., Ramsey et al. 2002; Shipley 2000). In some cases, it is also possible to accurately infer the existence of new unobserved variables that are common causes of the observed variables (Silva et al., 2003; Richardson & Spirtes, 2003).

Causal Bayes net representations and learning algorithms allow learners to accurately predict patterns of evidence from causal structure and to accurately learn causal structure from patterns of evidence. They constitute a kind of inductive causal logic, and a logic of causal discovery. It is possible to prove that only certain patterns of evidence will follow from particular causal structures, given the Markov, Intervention and Faithfulness assumptions, just as only certain conclusions follow from particular logical premises, given the axioms of logic.

From: mherskovits@psych.ucarcadia.arcadia.edu

To: brook_russell@turing.carnegietech.edu

3:15 AM .Aug 5 2003

Righto Brook ....

Well, quadruple countervailing causal prevention sounds just fascinating. I'm so glad I'm going to this conference now.

But thanks for the attachment. Actually, I think I might be getting the hang of these Bayes net things, even with all the formal stuff. (Though there's one thing about the math I still don't get, why do you Brits insist on making it plural?) They sound like something we know a lot about in Arcadia – vision. Not of course the political kind or the hallucination kind (although we know a lot about those, too) but the kind we study in psychophysics and perceptual psychology.

The world out there is full of real three dimensional objects but our perceptual system just gets some distorted 2 dimensional retinal input. Still, the merest "sprog" as you would say, has the computational power to turn that input back into a 3-d representation of a table or a lamp without even thinking about it. And (ignoring the occasional illusion) those representations are accurate – they capture the truth about the spatial world.

In vision science we have "ideal observer" theories about how that happens – how any system, animal, human or robotic, sprog or Ph. D. could infer the structure of a three-dimensional world from two-dimensional data. Vision science tells us that the visual system implicitly assumes that there is a world of three-dimensional moving objects and then makes assumptions about how those objects lead to particular patterns on the retina. By making the further assumption that the retinal patterns were, in fact, produced by the objects in this way, the system can work backwards and infer the structure of objects from those patterns (see e.g., Palmer, 1999).

Your causal Bayes net inferences sound sort of like that. The visual system assumes that the patterns at the retina were produced by three-dimensional objects in a particular way and then uses those assumptions to infer the objects from the retinal patterns. Your causal Bayes nets assume that causal structure produced patterns of evidence and uses those assumptions to learn the structure from the evidence (your Causal Markov, intervention and Faithfulness assumptions). You guys seem to think that you're going to do the same thing for causality that the psychophysicists have done for vision: you're going to tell us how we could transform information about probabilities and interventions into accurate representations of the causal structure of the world.

So I guess if you're right, (and I'm not committing myself yet there) causal Bayes nets COULD give us a way of formally specifying accurate inductive causal inferences – just like "ideal observer" theories in vision provide a way of formally specifying accurate visual inferences and like logic provides a way of formally specifying accurate deductive inferences

But if that's right, then I have to say, Brook, the rest of your letter doesn't make a whole helluva lot of sense :) You seem to be under the bizarre impression that any knowledge you can't find in the Times Literary Supplement isn't really "knowledge". So I guess you think my "sprogs" can't see because they can't write an article on Fourier transforms.

But of course my sprogs see just as well as you and I do. And of course, sprogs can use vision to learn all sorts of new things about objects. In fact, they engage in perfectly sophisticated "maths" all the time – and if they can perform complex, implicit computations to support vision, they could, in principle, perform complex, implicit computational procedures to support causal inference.

And the thing is, your computers may or may not be able to solve this causal learning problem – but it's damn sure that my sprogs can do it. In fact, they might be the most powerful causal learning devices in the universe. Thirty years of work on the theory theory shows that children have abstract, coherent, representations of the

causal structure of the world. Those representations allow children to make predictions, perform interventions and even generate counterfactuals. As soon as they can talk, they even offer explanations of the world around them. And they seem to learn those causal structures from patterns of evidence.

Plus even the very smallest sprogs can combine information from observation and intervention. Little babies who learn a new skill – like reaching for objects – understand other people's actions on objects better than babies who don't have the skill. Jessica Sommerville will show you next week how giving babies "sticky mittens" and changing their own ability to act on the world, changes the babies' ability to understand the actions of others. And Andrew Meltzoff will show you something like the reverse: how babies take information they only observe and turn it into actions of their own. And sprogs do all sorts of other things: make good interventions, discriminate confounded and unconfounded interventions, reason about unobserved causes, learn complex causal structure … Laura Schulz, Tamar Kushnir, and that Gopnik woman whose name you like so much will show you all that on Saturday too. And when it comes to grown-ups, York Hagmayer, Steve Sloman, Dave Lagnado and Michael Waldmann, will show you that even those stats class undergraduates can make remarkably sophisticated inferences about both predictions and interventions.

Best of all, sprogs never do absolutely useless things like reason about quadruple causal prevention.

Anyway, I'm doing my part and attaching some fairly primitive stuff about the psychology of causal learning. And as you'll see even the best theoretical accounts we have don't really even start to capture the richness of what people, even very small people, can actually do. All the best,

Morgan

Attachment 2: The Psychology of Causal Learning for Nerds

The Piagetian account of causal reasoning. Research on children's causal reasoning, like research on cognitive development in general, was initiated by the work of Jean Piaget. Piaget believed that causal reasoning developed very gradually. Indeed, Piaget proposed no less than seventeen distinct stages of causal learning.

In particular, however, Piaget believed that children's reasoning from early to middle childhood was "precausal". It was characterized by a "confusion between psychological activity and physical mechanism" (1930). This conclusion was based chiefly on his investigation of children's explanations of natural phenomena. Piaget found that children's early explanations of physical events were artificialistic (meaning events were attributed to human intervention -- clouds move because we walk; the river flows because of boats) and animistic (meaning that physical events were

attributed to psychological intention – the string turns because it wants to unwind itself; 1929). According to Piaget's account, not until quite late in development were children able to provide a complete, functional account of a chain of causal events and reason accurately about intervening causal mechanisms.

Nativist and modular views of causal reasoning. Over the past several decades however – and with the development of new methods for assessing the cognitive abilities of infants and young children – considerable research has suggested that Piaget underestimated the causal reasoning abilities of young children. Both infants and adults seem to perceive causality when objects (like billiard balls) collide and launch one another (Michotte, 1962; Leslie & Keeble, 1987; Oakes & Cohen, 1990). Infants also seem to expect causal constraints on object motion, assuming that objects respect principles of support, containment, cohesion, continuity and contact (Baillargeon, Kotovsky, & Needham, 1995; Spelke, et al., 1992; Spelke, Katz, Purcell, Ehrlich, & Breinlinger 1994).

Moreover, contra Piaget, considerable evidence suggests that even babies appropriately distinguish psychological and physical causality. Specifically, infants seem to interpret human, but not mechanical, action, as goal-directed and self-initiated (Meltzoff, 1995; Woodward, 1998; Woodward, Phillips & Spelke, 1993). Thus for instance, babies expect physical objects to move through contact (Leslie &

Keeble, 1987; Oakes & Cohen, 1990) but do not expect the same of human agents (Woodward, et al., 1993); expect that an object will be entrained when grasped by a human hand but not by an inanimate object (Leslie, 1982; 1984), and treat the reach of a human hand, but not the trajectory of a metal claw, as goal-directed (Woodward, 1998). Furthermore, almost as soon as children can speak they offer causal explanations (at least of familiar, every-day events) that respect domain boundaries (Hickling & Wellman, 2001). Finally, preschoolers' predictions, causal judgments and counterfactual inferences are remarkably accurate across a wide range of tasks and content areas (e.g., Gelman & Wellman, 1991; Kalish, 1996; Flavell, et al., 1995; Gopnik & Wellman, 1994; Sobel, 2004).

In order to account for the early emergence of structured, coherent, causal knowledge, some psychologists have suggested that children's early causal representations might be largely innate rather than learned. Following Kant's conception of a priori causal knowledge (1787/1899), some researchers have proposed that children's early causal understanding might originate in domain-specific modules (Leslie & Keeble, 1987) or from innate concepts in core domains (Keil, 1995; Carey & Spelke, 1994; Spelke, et al., 1994). These researchers have suggested that children's causal knowledge might be accurate not because of general learning mechanisms de-

signed to infer structure from evidence but because of specialized mechanisms dedicated to relatively constrained information-processing tasks (Leslie, 1994).

It may be that infants' object concepts, their ability to distinguish objects from agents, and their perception of Michottean causality do indeed have an innate basis. However, there seems less reason to believe that children's abilities to reason broadly about the causes of human behavior, physical events and biological transformations are an outgrowth of domain-specific modules. In particular, modular, domain-specific accounts of causal reasoning do not seem to explain how we identify particular causal relations within a domain, how we make causal inferences that transcend domain boundaries (i.e., that physical causes can be responsible for psychological effects and vice versa), and why causal reasoning is sensitive to patterns of evidence. Nonetheless, the majority of post-Piagetian research on preschool children's causal reasoning has emphasized the centrality of substantive, domain-appropriate principles.

Domain-specific causal knowledge, causal mechanisms, and the "generative transmission" account. In particular, researchers have focused on the role that substantive concepts, like force and spatial contact, might play in constraining young children's inferences about physical causal events (e.g., Bullock, Gelman & Baillargeon, 1982; Leslie, 1984; Shultz, 1982). In an influential monograph on children's causal

reasoning, the psychologist Thomas Shultz distinguished between a statistical view of causal relations, in which the causal connection between events is determined by the covariation of cause and effect, and a causal mechanism view of causality, in which causation is understood "primarily in terms of generative transmission" of force and energy (1982). In a series of experiments, Shultz demonstrated that in their causal judgments, preschoolers privilege evidence for spatially continuous processes compatible with the transmission of energy, over evidence for covariation. Preschoolers inferred, for instance, that a tuning fork whose vibrations were not obstructed was more likely to produce a sound than a tuning fork whose vibrations were blocked, even when the effect immediately followed an intervention on the latter and followed the former only after a delay. Similarly, Bullock, Gelman & Baillargeon concluded, that the idea that "causes bring about their effects by transfer of causal impetus" is "central to the psychological definition of cause-effect relations" (1982). Consistent with this view, psychologists have shown 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).

Covariation accounts. However, the generative transmission view of causation, in particular, and domain-specific knowledge, in general, have played a rather

limited role in accounts of adult causal learning. Indeed, in the adult cognitive science literature, researchers have largely focused on the role of contingency and covariation in causal learning, as opposed to principles about mechanisms. Two accounts of causal learning have been particularly influential: associative learning or connectionist accounts, and Patricia Cheng's causal power theory,

Associative learning and connectionist accounts of causal learning. Although not all contingencies are causal, all causal relationships involve contingencies. There is a vast literature on contingency learning in both human and non-human animals and some researchers have proposed that mechanisms similar to those underlying contingency learning in operant and classical conditioning can account for human causal reasoning (Dickinson, Shanks & Evenden, 1984; Shanks et al. 1996; Shanks & Dickinson, 1987; Wasserman, et al., 1993.)

Instrumental and imitative learning. Thorndike found that cats could learn to escape from cages by trial and error and that with practice, the cats became faster at escaping. He described this as the Law of Effect: actions with positive consequences are likely to be repeated and actions with negative consequences, avoided (1911). A large body of research on learning subsequently elaborated the ways in which behavior could be shaped by reinforcing or punishing outcomes. Operant learning has been demonstrated in non-human animals ranging from pigeons to primates and

unsurprisingly, it has been demonstrated in human babies as well. Thus infants who learn, for instance, that kicking makes a mobile spin, will both repeat the behavior and remember it after significant delays (Rovee-Collier, 1987; Watson & Ramey, 1987). Instrumental learning – the ability to learn from the immediate consequence of one's own actions – seems to be an early development, both phylogenetically and onotogentically.

Importantly, human beings (if not uniquely among animals, then at least characteristically see Tomasello & Call, 1997) are able to learn, not just from the consequence of their own actions but also from the consequences of others' actions. Thus, for instance, nine-month-old babies who see an experimenter light up a toy by touching it with his head will spontaneously touch their own heads to the toy (Meltzoff, 1988). By 18 months infants will even recognize the goal of another's intervention and produce the completed action when they have seen only a failed attempt (Meltzoff, 1995). Such research suggests that very young children can learn the causal relation between human actions and the events that follow them. However, it does not explain how children learn causal relations when human action is not the causal variable (e.g., the causal relationship between two parts of a toy, the causal relationship between growth and food, and the causal relationship between mental states and behavior). Instrumental learning and learning from the direct outcome of

others' interventions does not seem to explain our ability to engage in non-egocentric causal reasoning about distal events.

Classical learning and the Rescorla-Wagner theory. Shortly after Thorndike formulated the Law of Effect, Pavlov famously discovered that an animal regularly exposed to a temporal contiguity between a conditioned stimulus (like a tone) and an unconditioned stimulus (like food) would learn to associate the two stimuli. When presented only with the conditioned stimulus, the animal would produce a response (e.g., salivating) normally elicited by the unconditioned stimulus (1903). This finding has also been replicated across species and ages; like instrumental learning, classical conditioning is an ontogenetically and phylogentically, early and robust development.

Rescorla modified Pavlov's theory to suggest that contingency, not just contiguity, was critical for learning (1967). That is, in order for learning to occur, cues have to be predictive: the probability of the effect given the cue must be greater than the probability of the effect in the absence of the cue. The Rescorla-Wagner theory (R-W theory, 1972) specified that learning occurred on a trial-by-trial basis and predicted that early trials would be more important to learning than later trials. In its simplest form, the R-W equation for associative learning is: $\Delta V = K(\lambda - \sum V)$ where ΔV is the change in the perceived strength of the association (e.g., the amount of learn-

ing that occurs on any given trial), K is a parameter between 0 and 1 reflecting the salience of the cue multiplied by the salience of the effect, λ is the association between cue and stimulus at asymptote and $\sum V$ is the sum of the associative strength on previous trials. Thus, the R-W theory predicts that the change in associative strength on any trial is proportional to the difference between the maximum possible associative strength between a cue and an outcome and the previous estimate of the strength of association. Thus the stronger the prior association, the less learning on any given trial.

The model can be applied to human causal learning by substituting causes for the conditioned stimulus and effects for the unconditioned stimulus. The associative strength between the two variables is then taken as indicating the causal connection between them. This equation successfully predicts findings in the animal learning literature such as blocking, overshadowing, and conditioned inhibition and many findings in the human contingency learning literature (Baker et al., 1989; Dickinson et al., 1984; Shanks, Holyoak & Medin, 1996; Wasserman, et al., 1993). The R-W rule, or generalizations of the rule, have often been implemented in connectionist networks aimed at explaining human causal learning (see e.g., Gluck & Bower, 1988; Shanks, 1990, Rogers and McLelland 2004).

However, there is substantial agreement that the R-W equation by itself does not adequately account for the psychology of human causal learning (see e.g., Cheng, 1997; Glymour, 2002; Gopnik et al., 2004, Waldmann, 1992, 1996, 2000). In fact, it may not even explain animal learning. The R-W account predicts neither learned irrelevancy –the fact that an animal first exposed to a cue without any reward or punishment has difficulty on later conditioning trials learning to associate the cue with an outcome, nor failures of extinction – the fact that an animal who has learned through operant conditioning to avoid a cue once associated with a punishment, retains the behavior in the presence of the cue long after the association has disappeared.

In the human case, Patricia Cheng demonstrated for instance, that the R-W approach fails to account for boundary conditions on causal inference (1997). When an effect always occurs (i.e., whether the candidate cause is present or not) the R-W equation predicts that we should conclude that the candidate generative cause is ineffective. In contrast, human reasoners believe that it the effect occurs at ceiling there is no way to determine the efficacy of a candidate cause. Similarly, if an effect never occurs, the R-W equation predicts that we should believe a candidate inhibitory cause is ineffective, whereas people believe that if the effect never occurs it is impossible to determine the strength of an inhibitory cause. Similarly, Waldmann

(1992, 1996, 2000) showed asymmetries in the predictive and diagnostic uses of causal information that were difficult to explain in associationist terms.

The R-W account also fails to explain a phenomenon known as "backward blocking" (Sobel, Tenenbuam & Gopnik, 2004). If two candidate causes, A and B, together produce an effect and it is also the case that A by itself is sufficient to produce the effect, human reasoners (including young children) are less likely to believe that B is a cause of the effect. However, since observing A by itself provides no new evidence about the association between B and the effect, the R-W rule predicts that our estimate of the causal strength of B should not change (although some researchers, e.g., Wasserman & Berglan, 1998, have suggested modifications to the R-W rule that do allow for this prediction).

In addition to those aspects of human causal reasoning that seem to contradict the predictions of the R-W model, there are many aspects of human causal learning that would require ad hoc modification of the R-W rule. The R-W model for instance, calculates the strength of every candidate cause separately, thus to judge the interaction of two causes it must treat the interaction as a "third" candidate cause (see Gopnik et al., 2004). Similarly, the R-W equation assumes that all the variables have already been categorized as "causes" or "effects" and then calculates the associative strength between each cause and each effect. However, the model cannot de-

termine whether variables *are* causes or effects (i.e., it cannot decide whether A causes B, B causes A, or neither). One might run the equation multiple times, sometimes with one variable as a cause and sometimes with the other, and then compare the relative strength of each pairing, but this is an ad hoc modification of the theory.

The Power Theory of Probabilistic Contrast. Patricia Cheng (1997) proposed an account of human causal learning that resolved some of the difficulties with the R-W account. Cheng proposed that people innately treat covariation as an index of causal power (an unobservable entity) and suggested that people reason about causes with respect to particular *focal sets*, a contextually determined set of events over which people compute contrasts in covariation.

Cheng uses probabilistic contrast (ΔP) as an index of covariation. Delta P is simply the difference between the probability of an effect given a candidate cause and in the absence of the candidate cause; formally, $\Delta P = P(e \mid c) - P(e \mid \sim c)$. However, in distinction from purely covariational accounts of causal reasoning, Cheng introduces the idea of causal power. Although we cannot know the real causal power of any variable (since causal power is a theoretical entity) we can estimate causal power by distinguishing between the probability of the effect in the presence of a candidate cause and the probability the effect in the presence of all

causes (known and unknown) alternative to the candidate cause. Cheng assumes A) that candidate causes and alternative causes influence the effect independently; B) that there are no unobserved common causes of the candidate cause and the effect (although the account can be generalized to relax this assumption, Glymour, 2001) and, C) that candidate causes are non-interactive (although Novick and Cheng have since modified the account to explain inferences about interactive causes; 2004).

The causal power of a candidate cause is not equivalent to either P (e | c) or ΔP because even when the candidate cause is present and the effect occurs, the effect could be due to alternative causes. However, if you assume that alternative causes occur independently of the candidate cause, then the probability of the effect when the candidate cause is present and all alternative causes are absent can be estimated as 1 – P (e | ~c). Thus generative causal power (pc) can be estimated as: pc = ΔP / (1- P (e | ~c)).

As this equation illustrates, when alternative causes are absent, ΔP will reflect the causal power of c. However, as P (e | ~c) increases, ΔP becomes an increasingly conservative estimate of causal power. The limiting case of course, is when the effect always occurs (whether c is present or not). In that case, the reasoner can no longer use covariation as an index of causation and the causal power of c is undefined. This explains both why ceiling effects are a boundary condition on causal inference

and why covariation is not, in general, equivalent to causation. A parallel account explains inferences about candidate inhibitory causes.

Although compelling as a psychological account of human causal learning, one weakness of the power PC account is that, like the R-W account, it assumes that variables in the world are already identified as "causes" or "effects". The power PC account does not explain how, in the absence of prior knowledge or temporal cues, people could use data to distinguish causes and effects (i.e., to infer whether A causes B or B causes A).

Put another way, both the R-W account and the power PC account are explanations of how people judge the *strength* of different causal variables. These theories do not explain how people make judgments about causal *structure*. Additionally, neither the R-W nor the power PC theory provide a unified account of how people might go from judgments about causes to inferences about the effects of interventions. Finally, both of these accounts assume that the candidate causes and effects are observed. Neither account explains how people might use observational data to infer the existence of unobserved causes.

From: brook_russell@turing.carnegietech.edu

To: mherskovits@psych.ucarcadia.arcadia.edu

My dear Morgan, Thank you for your letter and the attachment. Well, perhaps you are right that there is more similarity between our problems than one might at first think. Your description of the different positions in the psychology of causal learning is indeed very reminiscent of the classical positions in the philosophical literature – partly, I suppose, because, historically speaking this is where the psychological positions ultimately come from. In philosophy accounts of causation have been similarly divided. Some accounts, like those of Dowe or Salmon, stress "mechanism" and "transmission". Much like your Shultz they argue that causation involves the spatio-temporal transmission of some sort of "mark" or "impetus" from cause to effect. Since Hume, the alternative account, usually phrased in skeptical terms, has been that causation just amounts to covariation – sounding rather like your associationists two centuries later. As, Bertrand Russell put it: "The law of causality, I believe, like much that passes muster among philosophers, is a relic of a bygone age, surviving, like the monarchy, only because it is erroneously supposed to do no harm".

But you see recently, and in tandem with all the new maths I told you about in that attachment, there's been a new way of thinking about causation in philosophy. Philosophers increasingly think about causation in relation to intervention: in terms that would suit your sprogs – if X causes Y, then if you wiggled X ,Y would

wiggle too. Jim Woodward will tell you all about it on Saturday and Chris Hitchcock will show you how it helps explain even those cases of quadruple countervailing prevention you find so amusing. And John Campbell will tell you how it applies to even the kind of causation your particular brand of scientist deals in – the psychological kind.

But here is the really important and I must confess, somewhat against my will, even intriguing thing about your letter. The unsolved problems you describe in the psychology of causal learning – the things you say your sprogs are so good at doing and the theories are so bad at explaining– well they're just the sort of things that the interventionist/causal Bayes net account seems well, destined for.

My learning algorithms, like your sprogs, can infer causal structure rather than just strength, they can appropriately combine information from interventions and observations, and distinguish appropriately between them, and they can even infer unobserved variables from evidence. So if the two actually were conjoined....

As ever,

Brook

P.S. Oh and, by the way, there seems to be a defect in your word-processing program. In several places where a full stop is clearly intended it seems to transmit a colon or semicolon followed by a right parenthesis instead: quite mysterious.

References

Ahn, W.-k., Kalish, C. W., Medin, D. L., Gelman, S. A., Luhmann, C., Atran, S., et al. (2001). Why essences are essential in the psychology of concepts. *Cognition, 82*(1), 59-69.

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.

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., & Spelke, E. S. (1994). Domain-specific knowledge and conceptual change. In L. A. Hirschfeld & S. A. Gelman (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* (pp. 169-200). New York, NY, US: Cambridge University Press.

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.

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).

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

Gluck, M., & Bower, G. H. (1988). Evaluating an adaptive network model of human learning. *Journal of Memory and Language, 27*(2), 166-195.

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

Glymour, C., & Cooper, G. F. (1999). *Computation, causation, and discovery*. Cambridge, MA: MIT/AAAI Press.

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., & 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.

Hickling, A. K., & Wellman, H. M. (2001). The emergence of children's causal explanations and theories: Evidence from everyday conversation. *Developmental Psychology, 37*(5), 668-684.

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

Kant, I. (1899). *Critique of Pure Reason (J. Meiklejohn, Trans.)*. New York: The Colonial Press. (Original work published 1787).

Keil, F. C. (1995). The growth of causal understandings of natural kinds. 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. 234-267). New York, NY, US: Clarendon Press/Oxford University Press.

Leslie, A. M. (1984). Infant perception of a manual pick-up event. *British Journal of Developmental Psychology, 2*(1), 19-32.

Leslie, A. M. (1982). The perception of causality in infants. *Perception, 11*, 173-186.

Leslie, A. M. (1994). ToMM, ToBy, and Agency: Core architecture and domain specificity. In L. A. Hirschfeld, S. A. Gelman (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. 119-148). New York, NY, US: Cambridge University Press.

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

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

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

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

Michotte, A. E. (1962). Causalite, permanence et realite phenomenales; etudes de psychologie experimentale. Louvain: Publications universitaires

Novick, L. R., & Cheng, P. W. (2004). Assessing interactive causal influence. *Psychological Review, 111*(2), 455-485.

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

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

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

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

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

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

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

Richardson, T. & Spirtes, P. (2003). Causal inference via ancestral graph models. In Green, P. Hjort N., & Richardson, S. (eds.) *Highly structured stochastic systems.* Oxford, Oxford University Press.

Rogers, T., & McLelland, J. (2004). *Semantic cognition: A parallel distributed approach.* Cambridge: MIT Press.

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

Schultz, T. R., Pardo, S., & Altmann, E. (1982). Young children's use of transitive inference in causal chains. *British Journal of Psychology, 72*(2), 235-241.

Shanks, D. R. (1990). Connectionism and the learning of probabilistic concepts. *Quarterly Journal of Experimental Psychology: Human experimental psychology, 42*(209-237).

Shanks, D. R., Holyoak, K., & Medin, D. L. (1996). *Causal learning*. San Diego, CA: Academic Press.

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.

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

Silva, R., Scheines, R. Glymour. C. and Spirtes., P. (2003). Learning measurement models for unobserved variables. *Proceedings of the 18th Conference on Uncertainty in Aritificial Intelligence,* AAAI Press.

Sobel, D. M. (2004). Exploring the coherence of young children's explanatory abilities: Evidence from generating counterfactuals. *British Journal of Developmental Psychology, 22*, 37-58.

Sobel, D. M., Tenebaum, J. & Gopnik, A. (2004). Children's causal inferences from indirect evidence: Backwards blocking and Bayesian reasoning in preschoolers. *Cognitive Science, 28*(3).

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.

Thorndike, E. L. (2000). *Animal intelligence: Experimental studies*. New Brunswick, NJ: Transaction Publishers. (Original work published 1911).

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

Waldmann, M. R, (1996). Knowledge-based causal induction. In

Shanks, D. R; Holyoak, K. et al. (Eds.) *Causal learning* (pp. 47-88). San Diego, CA, US: Academic Press.

Waldmann, M. R. (2000). Competition among causes but not effects in predictive and diagnostic learning. *Journal of Experimental Psychology: Learning, Memory, & Cognition.* Vol 26(1), Jan 2000, pp. 53-76

Waldmann, M. R; Holyoak, K. J, (1992). Predictive and diagnostic learning within causal models: Asymmetries in cue competition. *Journal of Experimental Psychology: General. Vol 121(2),* pp. 222-236.

Wasserman, E. A., & Berglan, L. R. (1998). Backward blocking and recovery from overshadowing in human causal judgment: The role of within compound associations. *Quarterly Journal of Experimental Psychology: Comparative and Physiological Psychology, 51*(2), 121-138.

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. (1972). Reactions to response-contingent stimulation in early infancy. *Merrill-Palmer Quarterly, 18*(3), 219-227.

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.* Paper presented at the Fifteenth Annual Meeting of the Cognitive Science Society.

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

Introduction to Part 2 – Causal Learning and Probability

From: brook_russell@turing.carnegietech.edu

To: mherskovits@psych.ucarcadia.arcadia.edu

Dear Morgan,

What an amazing weekend! Its completely changed my mind about the value of your kind of psychology and I simply cannot wait for the next workshop. My head is buzzing with questions and thoughts about human causal learning, and I'm so longing for the answers the next workshop seems all set to provide.

There is the whole question of probability, for example. I am struck that so many of those brilliant examples of causal learning you and the rest of them described, especially with the sprogs, seem restricted to deterministic contexts, in which causes always follow effects. Of course, causal Bayes nets can be applied to such contexts. But the canonical application of the systems is to cases involving what may be quite complicated systems of conditional probabilities. And while Bayes nets can be applied to deterministic systems such systems raise special problems for learning. Often they result in violations of Faithfulness But I'm sure that Thomas Richardson and Clark Glymour can take care of that if anyone can.

Along the same lines, I do still have that query about whether human beings of any age are really capable of calculating probabilities. Didn't one of your psychologists recently get a Nobel Prize for showing how bad even sophisticated adults were at probabilistic reasoning?

And I am curious also about the question of classification and categorization. The Bayes net formalism depends on the idea that variables are specified beforehand. That is, we already have some sense of how a particular event fits into a category – how a particular token is a member of a type as we say in philosophy - before we do any causal inference at all. But it appears that people often categorize objects precisely according to their causal powers. Could the formalism be applied to answer these questions too?

But I'm sure this next workshop will answer all this and more. All the best, Brook

From: mherskovits@psych.ucarcadia.arcadia.edu

To: brook_russell@turing.carnegietech.edu

Hi Brook, Well I have to say I feel the same way. Remember that quote from Gopnik I sent in that first letter? Well, I guess it's too soon to say for sure, of course, but this does seem awfully like the real thing. For once, a computational set of ideas really does seem to make contact with the things we care about in psychology. And even better, it gives us psychologists all sorts of new work to do. I can think of a zillion experiments to do to test the ideas already

And you know, I think a lot of your questions are going to be answered at the next workshop. It's true that people are just awful at explicitly representing prob-

abilities. But David Sobel and Natasha Kirkham will show you that even tiny babies, as young as eight months old, already can do some kinds of statistical reasoning, in fact, they already seem to use a kind of "screening-off". And Dave Lagnado and his colleagues will show you that adults can use probabilities to infer causation when they're combined with the right kinds of other cues, while Richard Scheines will show that even those undergraduate statistics students are, well, a lot smarter than they look.

As for categorization that's an interesting one. Because for a long time in psychology people noticed that "causal powers" seemed to play an important role in the way people categorized objects. In fact, one of the first areas of psychology where people talked about the theory theory was precisely in the domain of categorization. Adults' categories seemed to be based more on their theories of the deeper causal powers of objects -- their "essences" -- than on more superficial perceptual features. And David Danks and Bob Rheder will show how we can use the Bayes net formalism to make quite precise predictions about how ordinary folk will categorize objects.

Introduction to Part 3 – Causation, theories, and mechanisms

From: mherskovits@psych.ucarcadia.arcadia.edu

To: brook_russell@turing.carnegietech.edu

Brook,

I have to admit that I'm having second thoughts. Its true that its all very exciting still. But I worry that the gaps – the differences in background assumptions and interests and concerns – are just going to be too great to overcome.

In some ways the normative computational project seems like just the opposite of the psychological project. The computationalists, after all, are most interested in designing systems that can do just the things that human beings can't, like extracting causal structure from masses of correlational data all at one time. People, and children especially, seem to do things very differently. As Thomas Richardson pointed out, like scientists themselves, children can do experiments and they can do them repeatedly. And they make inferences from very small samples instead of the enormous data-bases that the computer systems operate on. But, ironically, we don't seem to have very good computational accounts of precisely how this sort of experimentation leads to accurate causal conclusions.

And there's another thing that's bothering me. Remember the whole point of this in the first place was to explain the nature and development of our intuitive theories. But I think I'm losing the connection between this sort of general causal learning and theory-formation. For instance, one of the main functions of intuitive

theories is to provide explanations – I know Henry Wellman has tons of terrific data about that. But there doesn't seem to be anything in the formalisms that corresponds to explanation. And theories also seem to constrain the kind of causal inferences we can make. When we have a theory the theoretical laws we formulate and the assumptions we make influence the very way we interpret the evidence. But again there doesn't really seem to be a good place for that kind of top-down effect of prior knowledge in the formalism.

And there's one more thing. I can't seem to get rid of this nagging sense that all those intuitions about mechanisms must come from somewhere – they must play some role or other. But its not at all clear just what that role is or how ideas about mechanism fit with causal Bayes nets. Maybe "mechanisms" are just more and more little arrows connecting the variables. But I think there must be more to it than that.

Anyway, maybe this is just a temporary let-down. But I thought I should let you know and see what you think about it. All the best, Morgan

From: brook_russell@turing.carnegietech.edu

To: mherskovits@psych.ucarcadia.arcadia.edu

Dear Morgan, I have to confess I've been having some of the same doubts myself. But you know I was looking at the papers for this last workshop and I do think that

there may be some answers there. Woo-Kyung Ahn and Michael Strevens are both going to talk about how we can integrate ideas about mechanism and causal structure and intervention. And Clark Glymour gives an example of how you could adjust your assumptions about causality to apply these ideas to the particular domain of social relations. And then Josh Tenenbaum, Tom Griffiths, and Sourab Niyogi are going to discuss ideas about representing theories and showing how the prior knowledge encoded in those theories can shape inferences.

But anyway, surely the measure of any relationship isn't just the initial excitement but the potential for long-term productivity. If this one works it won't be because all the problems are solved but because we have a succession of ideas, thoughts, experiments, discoveries, each unfolding from the one before. And we can gain strength from both the similarities and the differences if we change and evolve together. So let's see how it goes at the workshop and be patient and think hard and hope for the best. Which would, after all, be very good indeed. Brook