Philosophical Psychology
Vol. 18, No. 2, April 2005, pp. 259–272

Routledge
Taylor & Francis Group

The Supposed Competition between Theories of Human Causal Inference

David Danks

Newsome ((2003). The debate between current versions of covariation and mechanism approaches to causal inference. Philosophical Psychology, 16, *87–107.) recently published a critical review of psychological theories of human causal inference. In that review, he characterized covariation and mechanism theories, the two dominant theory types, as competing, and offered possible ways to integrate them. I argue that Newsome has misunderstood the theoretical landscape, and that covariation and mechanism theories do not directly conflict. Rather, they rely on distinct sets of reliable indicators of causation, and focus on different types of causation (type vs. token). There are certainly debates in the research field, but the theoretical landscape is not as fractured as Newsome suggests, and a potential unifying framework has already emerged using causal Bayes nets. Philosophical work on causal epistemology matters for psychologists, but not in the way Newsome suggests.*

1. Introduction

The vast majority of our decisions are influenced—at least in part—by our beliefs about the causal structure of the world. Thus, an obvious psychological problem is to determine the source(s) of those beliefs, including both prior knowledge and *de novo* learning. The last 20 years have witnessed a (relative) explosion of psychological research on this problem, resulting in a range of theories and experiments. Newsome (2003) recently undertook a much-needed critical review of some of those theories. Unfortunately, his review is marred by (i) the mistaken belief that the two current types of human causal inference theories (a) are the only active theory types and (b) are direct competitors; (ii) overly ambitious criteria for determining the value of a theory of causal inference; and (iii) a failure to discuss an extant theory that arguably performs the very unification for which he calls.

Correspondence to: David Danks, Department of Philosophy, 135 Baker Hall, Carnegie Mellon University, Pittsburgh, PA 15213, USA. Tel.: (412) 268-8047; Email: ddanks@cmu.edu

ISSN 0951-5089 (print)/ISSN 1465-394X (online)/05/020259-15 © 2005 Taylor & Francis Group Ltd
DOI: 10.1080/09515080500169371

Psychological theories of causal inference are typically placed into one of two camps, though the actual theoretical state of affairs is not this clean. (I will return to this point later.) Covariation theories model people's inferences of causal relationships between variables (types) when they are provided with statistical information, but are unable to use prior knowledge to any substantive degree. There are many different covariation theories; Newsome reviews a subset of them: specifically, Cheng and Novick's (1990, 1992) probabilistic contrast model, Cheng's (1997) power PC theory,[1] and Glymour's (1998, 2000) extension of the power PC theory to Bayes nets composed of noisy-OR and noisy-AND gates. Mechanism theories focus on inference of the particular causal relationship active in an event (token) given prior knowledge and possibly some statistical information. The only current mechanism theory, also reviewed by Newsome, was developed by Ahn and her collaborators (e.g. Ahn & Bailenson, 1996; Ahn, Kalish, Medin, & Gelman, 1995). Newsome raises difficulties and notes open problems for all of these theories (though he has essentially no discussion of the various empirical supports for each of the theories). He then argues that there are deep differences in the theoretical commitments of the two theory types. Despite the differences he finds, Newsome argues that the two theory types could be usefully integrated into a single theory of human causal inference. Moreover, he concludes that this integration of psychological theories would be significantly advanced by attending to definitions and distinctions that emerge from philosophical work on the nature of causation and mechanisms.

Unfortunately, Newsome fundamentally misrepresents the theoretical landscape. Throughout his paper, he consistently describes the theories and makes arguments as though covariation and mechanism theories are direct competitors. That is, he seems to think that it is impossible that both mechanism and covariation theories could be correct (in any substantive sense). This misunderstanding is particularly surprising, since Newsome liberally cites Glymour (1998, p. 39), who claimed in his abstract that mechanism versus covariation is a 'false and confused dichotomy'. Moreover, Newsome attempts to integrate the two theory types, which is only sensible if they do not directly contradict each other. Perhaps because he does not fully recognize this, his integration is of the 'mix-and-match' variety: he incorporates only some elements from covariation and others from mechanism theories, and leaves much out. It is not a true integration.

In the following sections, I argue that covariation and mechanism theories of human causal inference need not be direct competitors. In particular, the third section of this paper explores the interaction between metaphysical and epistemological theories of causation to try to determine whether we can conclude anything substantive from the fact that human causal inference in the two psychological theory types is based on different indicators of causation. Before diving into that issue, however, the second section attempts to clarify the target of theories of human causal inference: what are they trying to explain? In the fourth section, I describe a psychological theory (or family of theories) that has emerged in the past five years that argues that people use (something like) causal Bayes nets to represent causal beliefs. Covariation and mechanism theories are quite easily unified in this framework, particularly when we examine what is really required theoretically to explain the mechanism theory experiments. Contrary to Newsome's advice, I argue that there is no reason for psychologists currently working in this field to be deeply focused on philosophical accounts of 'mechanism'.

Of course, all of this is not to suggest that there are no real debates in the field of human causal inference. There are substantial debates between covariation theorists about the exact processes people use to learn from statistical information. There are disagreements about the amount of prior knowledge brought to bear in the various experimental settings. There are theoretical questions about the roles of unconscious processing versus deductive reasoning. There is emerging evidence that people use a range of strategies for causal learning, but these individual differences are relatively unexplored. But all of these questions are essentially orthogonal to the covariation versus mechanism debate on which Newsome focuses, and which is the focus here.

2. What Must a Theory of Causal Inference Explain?

Before focusing on whether the two types of theories are competitors, we should be clearer about exactly what they are intended to explain. That being said, I will not offer an explicit positive account of the domain of human causal inference theories, primarily because there does not seem to be any such single, sharply defined domain. Rather, there are a variety of 'target questions', as will become clearer in subsequent sections. This section thus argues only that there are some problems a theory of human causal inference should *not* be expected to explain at this stage in research.

Specifically, Newsome argues that covariation theories do 'not provide an adequate account for one important issue: how do cognizers identify candidate causes and relevant evidence from the indefinitely large pool of possible representations?' (p. 93).[2] That is, since covariation theories operate on some small set of cause/effect variables and there are infinitely many different variables we can use to represent the world, covariation theories are hopelessly underspecified and do not (in their current form) apply to actual, computationally bounded, people. This worry seems to be Newsome's primary (only?) conceptual objection to covariation theories. Covariation theories would apparently be good models for human causal inference in the absence of prior knowledge, if only they could explain how people determine what variables exist in the world, as well as which are potentially causally relevant in each particular domain.

The general problem to which Newsome alludes—that of variable definition and identification—is a hard, interesting problem. Moreover, this problem arises not just for theories in cognitive psychology, but also for algorithms in a variety of data mining and machine learning contexts, since most machine-learning algorithms presuppose some specification of the data in terms of variables. And some covariation theorists recognize explicitly that their theories fall short of solving the variable definition problem. For example, Cheng's (1997, p. 370) exposition of the power PC theory noted that her theory concerns causal inference 'when candidate causes and effects are clearly defined'. Alternately, there are various suggestions for solving this problem that Newsome overlooks: Glymour (2000) suggests, on computational and statistical grounds, several heuristics for forming new variables for causal inference; more famously, Eleanor Rosch and many others have given psychological evidence for different levels of feature selection (i.e. variable definition), although this work has not been well integrated into studies of causal inference (e.g., Rosch & Mervis, 1975).

However, the gap Newsome identifies—though present—does not seem to be one we should expect a theory of causal inference to close. Arguing that covariation theories are somehow incomplete because they do not solve this problem is analogous to claiming that a psychological theory of decision making is incomplete unless it provides an account of the perceptual processes by which the decision maker learns about her environment. No one would deny that decision making almost always requires obtaining some sort of information from the environment, but we do not criticize a psychological theory of decision making for taking perceptual information as a given. Of course, we would like to have both accounts, but it seems reasonable to tackle these problems separately, at least at first. Our scientific theories are inevitably bounded in scope, and so each must take some kind of information or representation as a well-defined input. For the moment, research on human causal inference takes well-defined variables as input. Newsome seems to expect too much when he chastises the covariation theorists for bracketing off the variable definition problem, particularly since the psychologists explicitly state what features of the problem are assumed to be well defined.

But perhaps I have misunderstood Newsome's criticism. The quoted passage suggested two distinct problems: parsing the world into variables, and labeling some of those variables as potentially causally relevant to each other. The above discussion argued that the former problem can correctly be bracketed off by covariation theorists for now. Perhaps Newsome's concern is really the latter problem: for which cluster of variables should we do causal inference (whatever that might be)? For example, he also says that '[causal] induction requires constraints because of computational complexity' (p. 94). This passage suggests a concern that, even if we have a solution to the variable definition problem, there will still be many more variables in a particular setting than could possibly be simultaneously considered by a computationally bounded person. Given a list of descriptive variables for any particular situation, how does the causal learner decide which variables to consider? Rather than requiring a covariation theory to explain variable definition, Newsome might just be asking for a model of the labeling of some subset of variables as potentially causally relevant.

There are several different responses a covariation theorist could make at this point. One response might plausibly be to respond that this problem can also legitimately be bracketed off by a covariation theory. Given the extreme diversity of possible situations, this worry threatens to grow into essentially one version of the frame problem: given an arbitrary situation, are there (logical, computational) rules for determining which variables are relevant—causally relevant, in this case?

A covariation theorist could reasonably maintain that solving the 'frame problem' falls outside of the domain of her theory.

A different response to this problem would be to try to develop exactly the sort of theory Newsome requests. One type of theory could argue that people try to learn, one by one, the variables that are causally connected in some domain, and then integrate the various distinct learned relationships. There is a normative theory for integrating separately learned causal relationships into a single causal structure (Danks, 2002), as well as psychological evidence that people do integrate causal relationships in relatively sensible ways (Hagmeyer & Waldmann, submitted). A different route would be to attempt to give a theory directly explaining which variables are labeled as potentially causally relevant. Gopnik, Glymour, Sobel, Schulz, Kushnir, & Danks (2004) have an extended discussion of different possible methods for picking out relevant variables for their theory of causal inference as Bayes net learning. Lien and Cheng (2000) provide some data about which properties of novel materials are picked out as causal. Alternately, several theories have been offered to explain how people determine which properties of an object are 'projectible' (to use the philosophical jargon) to other objects (e.g. the neural networks/neuroscience model offered by McClelland & Rogers, 2003). These theories could provide the resources to determine which variables in a particular case are viewed as 'potentially causally relevant.'

There is one final interpretation of Newsome's worry: namely, that he is concerned that most covariation theories assume a prior division of the relevant variables into potential causes and the effect. That is, most extant covariation theories assume more than just a small set of well-defined, potentially causally relevant variables. They also assume that we already know which of those variables are the potential causes, and which is the putative effect. The various theories leave open the exact process by which the variable ordering is established—perhaps prior knowledge, perhaps temporal information, perhaps some process based on the first few instances. This assumption is a serious, well-justified concern about covariation theories, since it seemingly forces causal learning into a highly artificial framework. We often are confronted with causal learning situations in which we do not know what is cause and what is effect, and these covariation theories simply do not apply to those situations. That being said, I will later advocate a different theory of causal learning—namely, causal Bayes nets—that does not require this *a priori* division, and so I will not dwell further on the problem here.

We might also ask whether mechanism theories—which also attempt to explain causal inference—have any machinery to explain either variable labeling or cause/effect separation. Newsome identifies mechanism theories as supposing that people 'might use their understanding of the causal structure of the world and their knowledge of potential mechanisms to (1) identify candidate causes . . . , and (2) identify relevant evidence' (p. 96). The only reason Newsome does not make these resources available to a covariation theorist is that he seems to think that the two types of theories are mutually exclusive. If the two types of theories are not direct competitors, then there is no reason to deny the covariation theorist the resources available to the mechanism theorist (in particular, prior domain knowledge and analogical reasoning). If they are instead compatible, then the problem of variable labeling is either a problem for every theory of causal inference, or none of them. We are thus led to ask whether the two theory types really compete.

3. The Compatibility of Covariation and Mechanism Theories

We can immediately point out one way in which the two theory types do *not* currently compete: covariation theories focus on type causation, and mechanism theories on token causation (though this separation is largely contingent—see below). Covariation theories talk about learning 'Smoking causes cancer', and mechanism theories talk about whether 'John's smoking caused his cancer'. That being said, Newsome quite correctly points out that we cannot dismiss the possibility of competition this easily. The theory types focus on different aspects of causal inference, but we should naturally expect them to extend their reach into each other's sub-domains, producing covariation theories of token causation and mechanism theories of type causation. If, as Newsome argues, they are radically different theory types, or have substantively different theoretical commitments, then we can anticipate that problems will emerge when these extensions happen. We should thus attempt to identify and reconcile those differences now, rather than later. In fact, these extensions are already occurring, at least at the normative level. Pearl (2000), Halpern and Pearl (2001a, b), Hitchcock (2001), and Glymour and Wimberly (in press) all offer accounts of token causation based on Bayes nets, the formalism to which Glymour (1998, 2000) extended the power PC theory. Psychological investigation of these accounts has only recently begun (e.g. Sloman & Lagnado, 2002).

The situation is roughly analogous to a hypothetical debate between an exemplar theory of categorization and a feature-based theory of concept learning. On one level, the theories focus on different phenomena (categorization judgments in the former, concept acquisition in the latter), but the two theories also make fundamentally different assumptions about the ways in which people learn and think about their world. Thus, any natural extensions of the two theories will directly compete, and so we can try to design 'critical experiments' now whose outcome might conclusively rule in favor of one or the other type of theory.

If the two theory types make radically different theoretical assumptions, then it seems reasonable to worry about their mutual consistency. Newsome argues that they have different theoretical assumptions because they are based (perhaps implicitly) on different theories of causation. This latter claim seems to be based on roughly the following line of reasoning. Philosophical accounts of the nature of causal relationships have tended to treat causation as a well-defined, unitary 'natural kind' whose precise specification is the object of debate (though see Skyrms, 1984, for a very different perspective). Regardless of the particular metaphysical account, the unitary nature of causation should lead to some relatively narrow, coherent set of reliable indicators of causation. These indicators need not be constitutive of causation, of course, but they should be reliable signals of it. It is also critical that people's beliefs about the causal relationships in our world be approximately correct if they are to plan, predict, take action, and so on. Therefore, we should expect that human cognition will pick up on this well-defined, unitary set of signals indicating causation, whatever it might be. Covariation and mechanism theories do, in fact, point towards different indicators (probabilistic relationships in covariation theories, some sort of 'transfer of power' in mechanism theories), and so—at the very least—must make different theoretical commitments.

Indeed, Newsome sometimes goes further and argues that the two types of psychological theories truly have different *metaphysical* commitments about causation. For example, he writes that 'covariation theorists assume that causal facts are reducible to non-causal facts about consistent patterns of covariation' (p. 99). This is simply false, and seriously misreads an explicit literature. The paradigmatic covariation theory for Newsome is Cheng's (1997) power PC theory, which was heavily influenced (unconsciously at the time, consciously now) by Cartwright's (1989) theory of causal capacities. Cartwright argues simultaneously that (i) causal relationships can be learned from statistical relationships in the world (given certain background knowledge or assumptions), but also (ii) causation is not reducible simply to those statistical relationships. No reduction of causation to probability is envisioned, and such a reduction has been explicitly denied by both Cartwright and Cheng.

In a similar vein, Newsome suggests that another point of difference between the covariation and mechanism theories is 'whether all the fundamental laws of nature are universal [i.e. mechanism theories], or at least some laws are irreducibly statistical [i.e. covariation theories]' (p. 97). But such remarks miss the very point of this psychological theorizing. These are theories of how people think about causation and make judgments and inferences about causation. Neither covariation nor mechanism theorists claim people are metaphysicians of any particular kind, nor need they. The covariation theorist can readily admit the existence of universal (deterministic) laws, while still thinking that people often draw causal conclusions from the probabilistic (due to lack of knowledge) phenomena we observe in the everyday world. And the mechanism theorist can allow that some fundamental laws are statistical, as long as there are still signals of the 'transfer of power.' The signals may not always occur, but that does not make them unreliable indicators of a causal connection.

Setting aside these confusions about the psychologists' metaphysical versus epistemological commitments, Newsome's 'competition position' assumes that causation is a well-defined 'natural kind', and concludes that there 'should' (in some sense) be some unique, narrow set of reliable indicators of causation that people exploit. However, this conclusion simply does not follow. Causation can be a single, well-defined type, and yet there might be multiple ways of learning about particular causal relationships. For example, even if a conserved/invariant quantity theory (as in Dowe, 1992, 2000) is correct, there could still be different reliable indicators of causation in different domains. Physical contact will be a crucial indicator for object collision (as classically demonstrated in Michotte, 1963), but physical contact might be irrelevant as an indicator for 'social causation' (e.g. my daughter's cry causing me to talk to her gently). Even though we might believe (under this theory of causation) that the full social causation story involves physical contact (of vibrating air molecules, released neurotransmitters, and so on), visible physical contact (or lack thereof) is not a reliable indicator of causation in this situation because the relevant physical contact is unobserved. Even if we accept the claim that causation—or even just our beliefs about causation—forms (or should form) a single 'natural kind', we cannot conclude that there must be a narrow set of reliable indicators of causation. If we cannot draw this conclusion, then there is no compelling justification for viewing the covariation and mechanism approaches as competitors.

In fact, there is some justification for drawing the opposite conclusion. As the above discussion indicates, it seems to be an empirical fact about our world that there actually are quite diverse signals of causation in different domains. And the existence of multiple reliable indicators does tell us something about the epistemology of causation: namely, that there is no reason to expect that people have only one causal inference algorithm. For example, there may be a (close-to-)hard-wired cognitive 'module' for making causal inferences from the Michotte-ean perception of object collisions, in addition to procedures for learning from data alone, from prior knowledge, and so on (for more on the idea of multiple psychological causal inference procedures, see Gopnik et al., 2004). And of more immediate relevance, it is entirely plausible—given the range of indicators of causation—that *both* the covariation and mechanism theories are correct, but that they are applied in different settings and domains. Or perhaps the covariation and mechanism theories are two different types of inference on a single representation of causal relationships. We now turn to explore this latter possibility.

4. Integrating Covariation and Mechanism Theories

If we recognize that covariation and mechanism theories are not necessarily competitors, then we can ask whether they can be incorporated into a single representational framework. In the past five years, there has been a substantial convergence of evidence from both cognitive and developmental psychology supporting the hypothesis that people represent (at least some of) their causal beliefs using a representation roughly like a Bayes net (e.g., Danks, Griffiths, & Tenenbaum, 2003; Glymour, 1998, 2000; Gopnik et al., 2004; Lagnado & Sloman, 2002, 2004; Sloman & Lagnado, 2002; Steyvers, Tenenbaum, Wagenmakers, & Blum, 2003; Tenenbaum & Griffiths, 2001, 2003; Tenenbaum & Niyogi, 2003; Waldmann & Martignon, 1998). If people actually do represent some of their causal beliefs using (something like) Bayes nets, then the unification of covariation and mechanism theories follows quite naturally, at least at the theory-type level.

Without going into substantial technical detail, a Bayes net is composed of a directed acyclic graph with a node for each observed variable, and a probability distribution over values in the nodes (see Neapolitan, 2003; Pearl, 2000; or Spirtes, Glymour, & Scheines, 1993, for more details). The crucial component for our purposes is the directed graph, since it encodes qualitative information about causal connections. Specifically, if the Bayes net is given a causal interpretation, then $X \rightarrow Y$ in the graph means 'X is a direct cause of Y relative to the variables in the graph'. (Only a weak metaphysical account of causation is needed here; a stronger, more complete explication is given in Woodward, 2003.) Causal Bayes nets are clearly not appropriate representations for all types of causal beliefs or knowledge. For example, they do not provide good models of the spread of diseases. They also do not easily model systems with feedback or continuously varying (in time) causal influence, unless we substantially modify the representation. However, for a wide range of domains and problems, causal Bayes nets have proven to provide useful, accurate representations of causal relationships. And the above-noted psychological research has all supported the hypothesis that people unconsciously use something like a Bayes net (or at least the graphical component of a Bayes net) to represent some of their causal beliefs.

There has also been substantial theoretical work over the past few years showing how covariation theories of causal inference fit into the Bayes net framework. Some covariation theories (specifically, the probabilistic contrast and power PC theories) have been shown to be maximum likelihood estimators of parameters in a fixed-graph Bayes net (Glymour, 1998; Tenenbaum & Griffiths, 2001). That is, these theories correspond to ones in which people assume that the world has the structure $C \rightarrow E \leftarrow A$ (where C is the potential cause, E is the effect, and A includes all other alternative causes), and then try to estimate the parameter ('causal power' or 'causal strength') associated with the $C \rightarrow E$ edge. The two psychological theories just correspond to different beliefs about how causes exert influence on their effects (i.e. different parameterizations of the Bayes net).

In addition to mapping existing covariation theories into the Bayes net framework, there are arguably covariation-type theories that are 'native' to Bayes nets. Recall that covariation theories assume that people are given (statistical) data about the world, but have no substantive prior knowledge to apply. There has been substantial work by statisticians and computer scientists using Bayes nets to try to learn causal relations in these situations; there are numerous algorithms for learning (as much as possible about) the graphical structure of a Bayes net from data about variables in the world. Recent experimental results suggest that people are—in covariation-type settings—potentially using these algorithms (or close variants) to infer causal structure, modeled as the graphical structure of a Bayes net (Danks et al., 2003; Steyvers et al., 2003; Tenenbaum & Griffiths, 2001, 2003). In fact, one prominent current debate in the psychological literature concerns the precise methods used to learn the causal (Bayes net) structure: Bayesian learning (Steyvers et al., 2003; Tenenbaum & Griffiths, 2001, 2003) versus constraint-based learning (Gopnik et al., 2004) versus top-down learning (Waldmann & Martignon, 1998; Lagnado & Sloman, 2004).

Integrating mechanism-type causal inference theories into the causal Bayes net framework is a bit trickier, since mechanism theories have been expressed more vaguely and less quantitatively than covariation theories. In general, the experimental results obtained by mechanism-type theorists can all be explained through (i) an account of the introduction of intervening variables, and (ii) some theory of token causation. That is, people explain causation in a particular event by trying to find some sequence of (causally) intervening variables that had 'causally active' values (in a token-causal sense) during the event. (I do not provide an extended defense of this claim here, but only note that I am not the first to make this observation—see Glymour, 1998.) In the causal Bayes net framework, this minimal theory translates into (i) addition of variables—possibly intervening ones—to our Bayes nets as we learn more about the world (see Danks, 2002 for a normative account of this process), and (ii) application of one (or more) of the existing accounts of token causation using causal Bayes nets. Hence, the framework of causal Bayes nets includes the computational resources to incorporate mechanism theories.

Of course, as noted above, this 'integration' is quite programmatic, largely because mechanism theories have typically not been quantitatively specified. Newsome advocates the use of philosophical work on 'mechanism' to help refine the psychological notion. For example, he suggests that people's understanding of a particular causal connection might initially be highly statistical (as in Salmon, 1984), but then shift to the entity/activity-based notion of Machamer, Darden, & Craver (2000) as they acquire more information about the various properties and processes involved in the connection. Little is known about changes in the properties of people's causal beliefs (e.g. statistical vs. activity based) as they acquire more information about a situation, and so this account may well be true. That being said, although the psychological and philosophical theories use the same word, the theoretical and experimental details are much more ambiguous about whether they are talking about the same concept, or even the same family of concepts. Consider just one example: Ahn et al.'s (1995, Experiment 1) result that people ask mechanism, rather than covariation, questions when trying to determine why some event occurred. Mechanism questions were picked out as those that (i) did not ask about other events; (ii) referred to novel factors (i.e. not previously identified); and (iii) could be viewed as a 'stand-alone' question. This criterion is clearly quite broad and covers a range of questions, and not necessarily those meeting some philosophical standard. The philosophical work on mechanism may prove to be useful to psychologists, but we will not know without substantially more research—both philosophical and psychological. Newsome also seems to advocate this type of research.

Perhaps most importantly, causal Bayes net theories provide a straightforward (though overly simplistic) way to think about a unification of the theory types: covariation theories explain how people *learn* causal relations, and mechanism theories explain how people *apply* causal relations. Covariation theories tell us how people learn, from statistical associations and background assumptions, causal relations that can be used both for prediction and for explanation in novel contexts. Mechanism theories attempt to describe how people use causal beliefs for explanation and prediction in particular cases. Mechanism theories assume that people have detailed prior beliefs about direct causes, causal intermediaries, generative versus preventive causes, and so on. Covariation theories attempt to explain the learning of exactly these sorts of beliefs. Covariation and mechanism theories are simply attempting to answer different questions about the world (specifically, learning vs. applying causal beliefs). And this distinction can be directly captured in the causal Bayes net framework as the difference between learning the underlying network structure (covariation theory) and performing various types of inference on some fixed causal structure (mechanism theory). Of course, this way of thinking about the issues is too simple: we can learn from mechanism information (e.g. when one presumed mechanism is explained away based on knowledge about other mechanisms), and we can apply covariation information (e.g. to obtain quantitative estimates of the impact of various decisions). Nevertheless, this description illustrates the ability of causal Bayes nets to explain both learning and application of causal beliefs.

The causal Bayes net framework also provides theoretical resources that are not clearly available in the other psychological theories. In particular, covariation theories do not explicitly model the distinction between observation and intervention, either in learning or prediction. That is, they assume that there is no difference between observing a variable's value and manipulating (i.e. forcing) the variable to have that particular variable.[3] The observation/intervention distinction can be introduced into these theories, but only in a somewhat *ad hoc* manner. It is unclear just what mechanism theories say about the observation/intervention distinction. In the causal Bayes net framework, on the other hand, this distinction is made quite clear and the different kinds of information can be differentially used in learning, prediction and decision making. Recent psychological work (e.g. Lagnado & Sloman, in press; Steyvers et al., 2003) has started investigated whether people seek out and use intervention information as predicted in the causal Bayes net framework.

Viewing theories of human causal inference through this causal Bayes net lens also helps to clarify the discussion in the second section about what the psychological theories should be expected to explain. The causal Bayes net representation offers no account of why variables are included or excluded. It assumes that there is some well-defined set of variables (though it does *not* assume a potential cause/effect labeling), and learning, updating and inference all take place on those variables. Through the course of learning, we might discover that some variable in the set is causally irrelevant to the problem at hand, but this learning does not explain why the variable appeared in the set in the first place. We might naturally wonder how we can (or should) include variables in a particular causal Bayes net, but this question is—at least at present—outside of the framework itself (though Spirtes and Scheines, 2004, point towards ways to postulate new variables). Thus, we should not expect any psychological theory that uses this framework to offer a solution to that problem. As noted earlier, Newsome simply expects too much of covariation theories; however, if this unification is correct, he also thinks too highly of mechanism theories, since he (mistakenly) believes that they have the resources to solve the variable labeling problem.

5. Conclusion

There are other minor points in the Newsome article about which one could quibble or complain. For example, he divides the covariation theories into associative and contingency models, without noting that some associative models actually make identical asymptotic predictions to some contingency models. The Rescorla–Wagner (1972) model, for example, is a quintessentially associative model that, in many cases, makes the same long-run predictions as Cheng and Novick's (1990, 1992) probabilistic contrast model, which is a paradigmatic contingency model (Cheng, 1997; Danks, 2003). A different associative model makes the same long-run predictions as the contingency-based power PC theory (Danks et al., 2003). There are certainly theoretical differences between associative and contingency theories; in particular, the former explain case-by-case changes in judgment, while the latter focus on asymptotic judgments. But the division between them is not necessarily an either/or choice. In many cases, a given associative theory does not conflict with a corresponding contingency theory, but rather offers one process explanation (out of many possible ones) of how people reach their long-term, stable, causal beliefs (as modeled by the contingency theory).

Newsome also makes some comments that suggest he believes that covariation theories assume that people explicitly calculate probabilities when trying to learn causal relationships from data. For example, Newsome's Table 1 states that 'people's conceptions of causality' are 'the probabilit[ies] with which causes influence the occurrence of the effect' (p. 94). And he later states that 'covariation theorists assume that people use their understanding of these constraining relations to infer the probability with which candidate causes produce or prevent target effects' (p. 98). Such claims are explicitly denied by covariation theorists. For example, Cheng (1997) understands people to be making ordinal judgments (such as '*A* is a stronger cause than *B*') that we theorists idealize as point-valued probabilities. No extant covariation theory requires that people be able to explicitly calculate probabilities, compute differences, or normalize values.

These latter complaints are relatively minor. The deeper issue throughout Newsome's critical review is his framing of the issue as covariation versus mechanism, when the real theoretical landscape is becoming 'covariation sometimes and for some problems, mechanism sometimes and for some problems'. The psychologist does not have to make an *a priori* choice between covariation and mechanism theories. The theory types attempt to explain different phenomena and different modes of inference, and so can usefully be explained by a single, unified, extant theory. There are many interesting, open problems in psychological research on causal learning, but 'covariation or mechanism?' is a false dilemma. In particular, we can potentially use the framework of causal Bayes nets to unify covariation and mechanism theories in a clear framework that both raises novel problems, and provides a deeper understanding of the relationship between the two theory types.

Notes

[1] Oddly, Newsome consistently refers to Cheng's theory as the 'PC power theory'.

[2] Unless otherwise indicated, this and all future page references are to Newsome (2003).

[3] This conflation is not as absurd as it might seem. If we know that *C* causes (or doesn't cause) *E* and there are no common causes of the two, then *E* has the same predicted value given either an observation or manipulation of *C*. The extant covariation theories assume (explicitly or implicitly) a particular causal structure in which the antecedent is satisfied.

References

Ahn, W., & Bailenson, J. (1996). Causal attribution as a search for underlying mechanisms: An explanation of the conjunction fallacy and the discounting principle. *Cognitive Psychology*, *31*, 82–123.

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

Cartwright, N. (1989). *Nature's capacities and their measurement*. Oxford, England: Oxford University Press.

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

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

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

Danks, D. (2002). Learning the causal structure of overlapping variable sets. In S. Lange, K. Satoh, & C. H. Smith (Eds.), *Discovery science: Proceedings of the 5th international conference* (pp. 178–191). Berlin: Springer-Verlag.

Danks, D. (2003). Equilibria of the Rescorla–Wagner model. *Journal of Mathematical Psychology*, *47*, 109–121.

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

Dowe, P. (1992). Process causality and asymmetry. *Erkenntnis*, *37*, 179–196.

Dowe, P. (2000). *Physical causation*. Cambridge, England: Cambridge University Press.

Glymour, C. (1998). Learning causes: Psychological explanations of causal explanation. *Minds and Machines*, *8*, 39–60.

Glymour, C. (2000). Bayes nets as psychological models. In F. C. Keil & R. A. Wilson (Eds.), *Explanation and cognition* (pp. 169–197). Cambridge, MA: MIT Press.

Glymour, C., & Wimberly, F. (in press). Actual causes and thought experiments. In M. O'Rourke (Ed.), *Explanation and causation*. Cambridge, MA: MIT Press.

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

Hagmayer, Y., & Waldmann, M. R. *Integrating fragments of causal models—implicit versus explicit sensitivity to structural implications*. Manuscript submitted for publication.

Halpern, J. Y., & Pearl, J. (2001a). Causes and explanations: A structural-model approach—Part I: Causes. In J. Breese & D. Koller (Eds.), *Uncertainty in artificial intelligence: Proceedings of the Seventeenth Conference* (pp. 194–202). San Francisco: Morgan Kaufmann.

Halpern, J. Y., & Pearl, J. (2001b). Causes and explanations: a structural-model approach—Part II: Explanations. In B. Nebel (Ed.), *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence* (pp. 27–34). San Francisco: Morgan Kaufmann.

Hitchcock, C. (2001). The intransitivity of causation revealed in equations and graphs. *Journal of Philosophy*, *98*, 273–299.

Lagnado, D., & Sloman, S. A. (2002). Learning causal structure. In W. D. Gray & C. D. Schunn (Eds.), *Proceedings of the Twenty-Fourth Annual Conference of the Cognitive Science Society* (pp. 560–565). Hillsdale, NJ: Erlbaum.

Lagnado, D., & Sloman, S. A. (2004). The advantage of timely intervention. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*, 856–876.

Lien, Y., & Cheng, P. (2000). Distinguishing genuine from spurious causes: A coherence hypothesis. *Cognitive Psychology*, *40*, 87–137.

Machamer, P., Darden, L., & Craver, C. F. (2000). Thinking about mechanisms. *Philosophy of Science*, *67*, 1–25.

McClelland, J. L., & Rogers, T. T. (2003). The parallel distributed processing approach to semantic cognition. *Nature Neuroscience*, *4*, 310–322.

Michotte, A. E. (1963). *The perception of causality*. New York: Basic Books. (Originally published in 1946).

Neapolitan, R. (2003). *Learning Bayesian networks*. Upper Saddle River, NJ: Prentice Hall.

Newsome, G. L. (2003). The debate between current versions of covariation and mechanism approaches to causal inference. *Philosophical Psychology*, *16*, 87–107.

Pearl, J. (2000). *Causality: models, reasoning, and inference*. Cambridge: Cambridge University Press.

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

Rosch, E. H., & Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology*, *7*, 573–605.

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

Skyrms, B. (1984). EPR: Lessons for metaphysics. In T. French, T. Uehling Jr, & H. Wettstein (Eds.), *Midwest studies in philosophy IX* (pp. 245–255). Minneapolis, MN: University of Minnesota Press.

Sloman, S. A., & Lagnado, D. (2002). Counterfactual undoing in deterministic causal reasoning. In W. D. Gray & C. D. Schunn (Eds.), *Proceedings of the Twenty-Fourth Annual Conference of the Cognitive Science Society* (pp. 828–833). Hillsdale, NJ: Erlbaum.

Spirtes, P., & Scheines, R. (2003). Causal inference of ambiguous manipulations. In S. D. Mitchell (Ed.), *Proceedings of the 2002 Biennial Meeting of the Philosophy of Science Association* (pp. 833–845). Chicago: University of Chicago Press.

Spirtes, P., Glymour, C., & Scheines, R. (1993). *Causation, prediction, and search* (2nd ed., 2001). Cambridge, MA: MIT Press.

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

Tenenbaum, J. B., & Griffiths, T. L. (2001). Structure learning in human causal induction. In T. Leen, T. Deitterich, & V. Tresp (Eds.), *Advances in neural information processing 13* (pp. 59–65). Cambridge, MA: MIT Press.

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

Tenenbaum, J. B., & Niyogi, S. (2003). Learning causal laws. In R. Alterman & D. Kirsh (Eds.), *Proceedings of the Twenty-Fifth Annual Conference of the Cognitive Science Society*. Mahwah, NJ: Erlbaum.

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

Woodward, J. (2003). *Making things happen: a theory of causal explanation*. Oxford, England: Oxford University Press.