Beyond covariation: Cues to causal structure

David A. Lagnado[1], Michael R. Waldmann[2], York Hagmayer[2], and Steven A. Sloman[3]

[1]University College London, UK

[2]University of Göttingen, Germany.

[3]Brown University, Providence, Rhode Island, USA.

To appear in Gopnik, A., & Schultz, L. (Eds.), Causal learning: Psychology, philosophy, and computation. In preparation.

Address for correspondence:

David A. Lagnado
Department of Psychology
University College London
Gower Street
London WC1E 6BT, UK
d.lagnado@ucl.ac.uk
Telephone: +44 (0) 20 7679 5389

Beyond covariation: Cues to causal structure

1. Introduction

Imagine a person with no causal knowledge, nor concept of cause and effect. They would be like one of Plato's cave dwellers – destined to watch the shifting shadows of sense experience, but know nothing about the reality that generates these patterns. Such ignorance would undermine that person's most fundamental cognitive abilities - to predict, control and explain the world around them. Fortunately we are not trapped in such a cave – we are able to interact with the world, and learn about its generative structure. How is this possible?

The general problem, tackled by philosophers and psychologists alike, is how people infer causality from their rich and multifarious experiences of the world. Not just the immediate causality of collisions between objects, but the less transparent causation of illness by disease, of birth through conception, of kingdoms won through battle. What are the general principles that the mind invokes in order to identify such causes and effects, and build up larger webs of causal links, so as to capture the complexities of physical and social systems?

2. Structure versus strength

When investigating causality a basic distinction can be made between *structure* and *strength*. The former concerns the qualitative causal relations that hold between variables – whether smoking causes lung cancer, aspirin cures headaches etc. The latter concerns the quantitative aspects of these relations – to what degree does smoking cause lung cancer, or aspirin alleviate headaches? This distinction is captured more formally in the causal Bayes net framework. The structure of a set of

variables is represented by the graph, the strength of these links captured in the parameterisation of the graph (the probabilities and conditional probabilities that, along with the graph itself, determine the probability distribution represented by the graph).

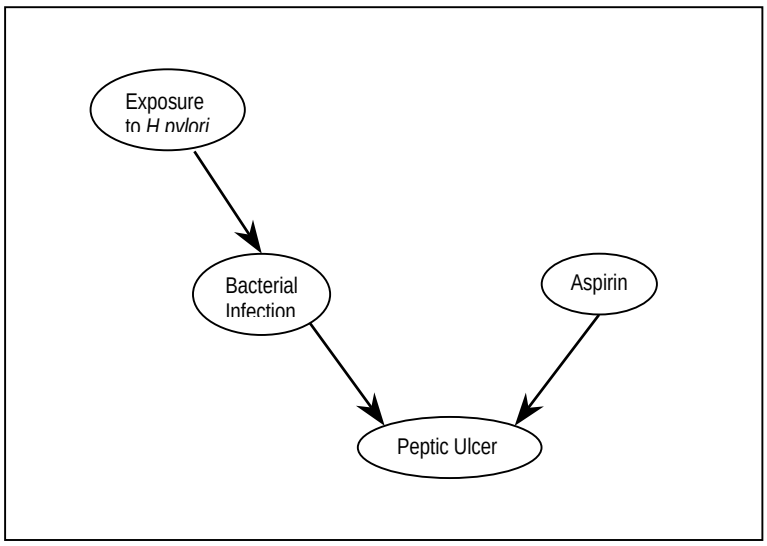

Figure 1. A simple Bayesian network representing the potential causes of peptic ulcers.

Conceptually, the question of structure is more basic than that of strength – one needs to know or assume the existence of a link before one can estimate its strength. This is reflected in many of the discovery algorithms used in AI, where there is an initial structure learning step prior to estimating the parameters of a graph (see Neapolitan, 2003). A natural conjecture is that this priority of structure over strength is likewise marked in human cognition (Pearl, 1988; Tenenbaum & Griffiths, 2001; Waldmann, 1996; Waldmann & Martignon, 1998).

This idea receives intuitive support. We often have knowledge about what causes what, but little idea about the strength of these relations. For example, most of us believe that smoking causes cancer, that exercise promotes health, that alcohol inhibits speed of reaction, but know little about the strengths of these relations.

Likewise in the case of learning, we seek to establish whether or not causal relations exist before trying to assess how strong they are. For example, in a recent medical scare in the UK, research has focused on whether the MMR vaccine causes autism, not on the degree of this relation. Indeed the lack of evidence in support of the link has pre-empted studies into how strong this relation might be.

The idea that causal cognition is grounded in qualitative relations has also influenced the development of computational models of causal inference. To motivate his structural account, Pearl (2000) argued that people encode stable aspects of their experiences in terms of qualitative causal relations. This inverts the traditional view that judgments about probabilistic relations are primary, and that causal relations are derived from them. Rather, 'if conditional independence judgments are by-products of stored causal relationships then tapping and representing those relationships directly would be a more natural and more reliable way of expressing what we know or believe about the world' (2000, p. 22).

Despite the apparent primacy of structure over strength, most research in causal learning has focused on how people estimate the strength of separate links. In a typical experiment variables are pre-sorted as potential causes and effects, and participants are asked to estimate the strength of these relations (e.g., Cheng, 1997; Shanks, 2004). This approach has generated a lot of data about how people use contingency information to estimate causal strength, and how these judgments are modulated by response format etc., but does not consider the question of how people learn about causal structure. Thus it fails to address an important (arguably the most fundamental) part of the learning process.

This neglect has had various repercussions. It has led to an over-estimation of the importance of statistical data at the expense of other key cues in causal learning.

For example, associative theories focus on learning mechanisms that encode the strength of covariation between cues and outcomes (e.g., Shanks & Dickinson, 1987), but they are insensitive to the important structural distinction between causes and effects. As a consequence they are incapable of distinguishing between associations that link spurious relations (e.g., barometer and storm) from true causal relations (atmospheric pressure and storm). More generally these models are incapable of distinguishing between direct and indirect causal relations, or covariations that are generated by hidden causal events (Waldmann, 1996; Waldmann & Hagmayer, in press).

Another shortcoming of this focus on strength is that it restricts attention to a small subset of causal structures (mainly common-effect models). For example, Power PC theory (Cheng, 1997) focuses on the assessment of causal strength based on covariation information. Although the main focus of the empirical studies lies in how people estimate causal power (see Buehner, Cheng, & Clifford, 2003), the theory clearly states that these power estimates are only valid under the assumption that the causal effect is generated by a common-effect structure with specific characteristics. The question of how people induce these models, which are a pre-requisite for the strength calculations, is neglected in this research. Moreover, people routinely deal with other complex structures (e.g., common-cause and chain models). The question of how people learn such structures, and how they combine simple structures into more complex ones, are clearly crucial to a proper understanding of causal cognition.

Furthermore, the focus on strength fails to give due weight to the importance of intervention (rather than passive observation), and to the temporal order of experienced events (over and above their temporal contiguity). Both of these factors are primarily cues to structure rather than strength, and there is growing evidence that

people readily use them (Gopnik et al., 2004; Lagnado & Sloman, 2004; Steyvers et al., 2003; Waldmann, 1996).

Even the traditional studies on strength estimation are open to re-evaluation in the light of the structure/strength distinction. Tenenbaum and Griffiths (2001) contend that participants in these studies are actually assessing the degree to which the evidence supports the existence of a causal link, rather than the strength of that link. More generally, they propose that people adopt a two-step procedure to learn about elemental causal relations, first inferring structure, and then estimating strength. Although decisive experiments have yet to be run, Griffiths and Tenenbaum (in press) support this claim through the re-interpretation of previous data sets and some novel experimental results.

The main moral to be drawn from these considerations is not that strength estimation has no place in causal learning, but that the role of structural inference has been neglected. By recognizing the central role it plays in both representation and learning, we can attain a clearer perspective on the nature of causal cognition.

3. Causal-model theory

Causal-model theory was a relatively early, qualitative attempt to capture the distinction between structure and strength (see Waldmann & Holyoak, 1992; Waldmann, Holyoak, & Fratianne, 1995; Waldmann, 1996; Waldmann, 2000, 2001; Waldmann & Martignon, 1998; Waldmann & Hagmayer, 2001; Hagmayer & Waldmann, 2002; see also Rehder, 2003a, b). According to this proposal causal induction is guided by top-down assumptions about the structure of causal models. These hypothetical causal models guide the processing of the learning input. The basic idea behind this approach is that we rarely encounter a causal learning situation in which we do not have some intuitions about basic causal features, such as whether

an event is a potential cause or effect. If, for example, the task is to press a button and observe a light (e.g., Wasserman, Chatlosh, & Neunaber, 1983), we may not know whether these events are causally related or not, but we assume that the button is a potential cause and the light is a potential effect. Once a hypothetical causal model is in place, we can start estimating causal strength by observing covariation information. The way covariation estimates are computed and interpreted is dependent on the assumed causal model (Waldmann & Hagmayer, 2001; Hagmayer & Waldmann, 2002).

The distinction between causal structure and causal strength raises the question of how assumptions about causal models are generated. Our working hypothesis is that people use a number of non-statistical cues to generate hypothetical causal models. We do not rule out the possibility that people occasionally induce causal structure on the basis of covariation information alone, but this seems rare in the world in which we live. Whenever people do not have clear assumptions about causal structure, causal reasoning easily falls prey to cognitive biases, such as confusing spurious with causal relations. In contrast, whenever people have hypothetical knowledge about causal structure they show a remarkable competence to tune this knowledge to the statistical relations in the learning input, and use this knowledge for predictions, diagnoses, and for planning actions.

4. Cues to causal structure

People are active agents immersed in a dynamic physical world. Not only do they experience events in a diversity of ways, but they experience a variety of relations between these events. Perhaps most significantly, they can also interact with the world, thereby creating new relations and disrupting old ones. The richness of

these experiences of the world affords people a variety of *cues* to its causal structure. Here is a partial list:

- Statistical relations
- Temporal order
- Intervention
- Prior knowledge

Following Einhorn and Hogarth (1986), we note that these cues are fallible, sometimes redundant (separate cues support the same conclusion), and at other times inconsistent (separate cues suggest opposing conclusions). These cues can be combined to construct and update causal models. For example, typical cases of intervention combine multiple cues -- proximity in space and time, temporal order, and covariation. This synergy explains the *power* of intervention as a route to causal knowledge. Cues are generally strongly correlated in natural environments -- causes tend to be nearby, prior to, and correlated with their effects

4.1 Statistical Covariation

Hume's analysis of causation has set the agenda for most contemporary theories of learning. These theories assume that causation cannot be perceived directly, and suppose that people infer it from the statistical patterns in what they can observe. The key idea is that people are exposed to patterns of data, such as the occurrence or non-occurrence of patterns of events, the presence or absence of features, or, more generally, the values of variables. From this body of data they extract certain statistical relations, upon which they base their causal judgments. There are various statistical relations that have been implicated in this process (Cheng, 1997; Glymour, 2001; Gopnik et al., 2004; Shanks, 2004). One of the simplest is the covariation between two events. For example, smoking increases the

probability of heart disease. The existence of a stable covariation between two events A and B is a good indication that *some* underlying causal relation exists, but by itself does not reveal whether A causes B, B causes A, or both are effects of a common cause. This highlights the incompleteness of any model of structure learning based solely on covariation detection.

The advent of Bayesian networks provides a more general framework to represent the statistical relations present in a body of observed data (Pearl, 1988). As well as representing straightforward (unconditional) relations between variables, they also represent conditional relations. In particular they represent relations of *conditional independence*. This holds whenever an intermediate variable (or set of variables) renders two other variables (or sets of variables) probabilistically independent. For example, the unconditional dependence between intravenous drug usage and AIDS is rendered independent conditional on HIV status. In other words, the probability that someone develops AIDS, given that they are HIV positive, is not affected by whether they contracted the virus through drug use (assuming that drug usage does not affect the passage from HIV infection to AIDS). Establishing the conditional independencies that hold in a body of data is a critical step in constructing an appropriate Bayesian network.

Recent work in statistics and AI forges a crucial link between statistical data and causal structure (Pearl, 2000; Spirtes et al., 1993). Given certain assumptions (e.g., the causal markov condition and faithfulness, see Woodward, this volume), they detail the patterns of dependencies that are associated with a given causal structure, and, conversely, the causal structures that can be inferred from a given pattern of dependencies. Based on this analysis a range of algorithms have been developed that can infer causal structure from large databases of statistical information. The success

of this computational work has prompted some to model human causal learning along similar lines (Glymour, 2001; Gopnik et al., 2004; see Section 7 for discussion).

Despite the sophistication of Bayesian networks, it is generally recognized that statistical data alone is insufficient for inferring a unique causal model. Even with the notion of conditional independence, a particular body of correlational data will typically be associated with several possible causal structures (termed Markov equivalent) rather than a unique model. For example, if it is known that A, B and C are all correlated (unconditionally dependent), and that A is conditionally independent of C given B, then there are three possible causal structures compatible with these relations (A →B →C, A←B→C, A ← B ←C). To narrow down these possibilities to just one requires some additional information. For instance, if one also knows that A occurs before B, then A →B →C is the only possible model.

This sets a theoretical limit on what can be inferred through correlation alone. At best statistical cues can narrow down the set of possible models to those that are Markov equivalent. There are also practical limitations. Even with just 3 variables there are a large number of correlations and conditional correlations to compute in order to determine viable causal models. And each of these relations requires a sizable amount of data before their individual reliability is established. Thus inferring possible causal models in a purely data-driven fashion involves a significant computational load. While this may be manageable by a powerful computer, it is less likely to be achievable by humans with limited processing and memory resources.

Indeed current evidence suggests that people are rather limited in their ability to learn structure from correlations alone, even to Markov equivalence. For example, Lagnado and Sloman (2004) presented subjects with probabilistic data generated by a three variable chain A →B→ C. In the absence of other cues (intervention, time order

etc.), most subjects failed to learn the correct structure or its Markov equivalents. This result holds up across several different learning paradigms (Lagnado & Sloman, in preparation; Steyvers et al., 2003; Sobel, 2003; Danks & McKenzie, forthcoming).

What people seem to find most difficult is establishing the appropriate conditional independence relations between sets of variables, and using this as a basis for inferences about causal structure. This is tricky because learners must track the concurrent changes in three different variables. They must determine whether the correlation between any pair of these variables is itself dependent on a third variable. For example, in Lagnado and Sloman (2004), participants had to figure out that (i) two different chemicals covaried with a given effect, and (ii) one of these chemicals was probabilistically independent of the effect conditional on the presence or absence of the other chemical. It is not surprising that most participants failed to work this out, and settled for a simpler (but incorrect) causal model.

The experiments of Steyvers et al. (2003) also demonstrated the difficulty of inducing structure from covariation data. In their experiments learners observed data about three mind-reading aliens. The task was to find out which of the three mind-readers can send messages (i.e., is a cause), and which can receive messages (i.e., is an effect). Generally, performance was better than chance but was still poor. For example, in Experiment 3 in which learners could select multiple models that are compatible with the data, only 20 percent of the choices were correct. This number may even overstate what people can do with covariation alone. In the experiments, learners were helped by the fact that the possible models were shown to them prior to learning. Thus, their learning was not purely data driven but was possibly aided by top-down constraints on possible models. Moreover, the parameters of the models were selected to make the statistical differences between the models quite salient. For

example, the pattern that all three mind-readers had the same thought was very likely when the common-cause model applied but was extremely unlikely under a common-effect model. Similarly, the pattern that only two aliens had the same thought was very likely under the common-effect model hypothesis but unlikely with chains or common-cause models. Under the assumption that people associate these different prototypic patterns (e.g., three mind readers with identical thoughts) with different causal structures (e.g., common-cause model), some participants might have solved the task by noticing the prevalence of one of the prototypic patterns. Additional cues further aided induction. As in Lagnado and Sloman (2004) performance improved when participants were given the opportunity to add an additional cue, interventions (see also Sobel, 2003; and Section 4.3).

In sum, there is very little evidence that people can compute the conditional dependencies necessary for inferring causal structure from statistical data alone without any further structural constraints. In contrast, when people have some prior intuitions about the structure of the causal model they are dealing with, learning data can be used to estimate parameters within the hypothetical model, or to select among alternative models (see also Waldmann, 1996; Waldmann & Hagmayer, 2001). Thus, the empirical evidence collected so far suggests that cues other than statistical covariation take precedence in the induction of structure before statistical patterns can meaningfully be processed. In the next section we show that the temporal order cue can override statistical covariation as a cue to causal structure.

4.2 Temporal Order

The temporal order in which events occur provides a fundamental cue to causal structure. Causes occur before (or possibly simultaneously with) their effects, so if one knows that event A occurs after event B, one can be sure that A is not a

cause of B. However, while the temporal order of events can be used to rule out potential causes, it does not provide a sufficient cue to rule them in. Just because events of type B reliably follow events of type A, it does not follow that A causes B. Their regular succession may be explained by a common cause C (e.g., heavy drinking first causes euphoria and only later causes sickness). Thus the temporal order of events is an imperfect cue to causal structure. This is compounded by the fact that we often do not have direct knowledge of the actual temporal order of events, but are restricted to inferring that order from the order in which we experience (receive information about) these events. In many situations the experienced order will reflect the true temporal order, but this is not guaranteed. Sometimes one learns about effects prior to learning about their causes. For example, the presence of a disease is typically learned about after experiencing the symptoms that it gives rise to (see Section 4.4 for further examples).

Despite its fallibility, temporal order will often yield a good cue to causal structure, especially if it is combined with other cues. Thus, if you know that A and B covary, and that they do not have a common cause, then discovering that A occurs before B tells you that A causes B and not vice versa. It is not surprising therefore that animals and humans readily use temporal order as a guide to causality. Most previous research, however, has focused on how the temporal delay between events influences judgments of causal strength, and paid less attention to how temporal order affects judgments of causal structure. The main findings have been that judged causal strength decreases with increased temporal delays (Shanks, Pearson & Dickinson, 1989), unless people have a good reason to expect a delay (e.g., through prior instructions or knowledge, see Buehner & May, 2002). This fits with the default assumption that the closer two events are in time, the more likely they are to be

causally related. In the absence of other information, this will be a useful guiding heuristic.

Temporal Order versus Statistical Data

Both temporal order and covariation information are typically available when people learn about a causal system. These sources can combine to give strong evidence in favor of a specific causal relation, and most psychological models of causal learning take these sources as basic inputs to the inference process. However, the two sources can also conflict. For example, consider a causal model in which C is a common cause of both A and B, and where B always occurs after A. The temporal order cue in this case is misleading, as it suggests that A is a cause of B. This misattribution will be particularly compelling if the learner is unaware of C. However, consider a learner who also knows about C. With sufficient exposure to the patterns of correlation of A, B and C they would have enough information to learn that A is probabilistically independent of B given C. Together with the knowledge that C occurs before both A and B the learner can infer that there is no causal link from A to B (without such temporal knowledge about C, they can only infer that A is not a direct cause of B, because the true model might be a chain A→C→B).

In this situation the learner has two conflicting sources of evidence about the causal relation between A and B – a temporal order cue that suggests that A causes B and (conditional) correlational information that there is no causal link from A to B. Here a learner must disregard the temporal order information and base their structural inference on the statistical data. However, it is not clear how easy it is for people to suppress the temporal order-based inference, especially when the statistical information is sparse. Indeed in two psychological studies Lagnado and Sloman

(2004, in preparation) show that people let the temporal order cue override contrary statistical data.

To explore the impact of temporal order cues on people's judgments about causal structure, Lagnado and Sloman (in preparation) constructed an experimental learning environment in which subjects used both temporal and statistical cues to infer causal structure. The underlying design was inspired by the fact that viruses (electronic or otherwise) present a clear example of how the temporal order in which information is received need not reflect the causal order in which events happen. This is because there can be considerable variability in the time of transmission of a virus from computer to computer, as well as variability in the time it takes for an infection to reveal itself. Indeed it is possible that a virus is received and transmitted by a computer before it reveals itself on that computer. For example, imagine that your office-mate's computer becomes infected with an email virus that crashes his computer. Twenty minutes later your computer crashes too. A natural reaction is to suppose that his computer transmitted the virus to you; but it is possible that your computer received the virus first, and then transmitted it to your office-mate. It just so happened that the virus subverted his computer more quickly than yours. In this case the temporal order in which the virus manifests itself (by crashing the computer) is not a reliable cue to the order in which the computers were infected.

In such situations, then, the order in which information is received about underlying events (e.g., the order in which viruses manifest themselves on computers in a network) does not necessarily mirror the underlying causal order (e.g., the order in which computers are infected). Temporal order is a fallible cue to causal structure. Moreover, there might be statistical information (e.g., the patterns of correlation between the manifestations of the viruses) which does provide a veridical cue to the

underlying structure. How do people combine these two sources of information, and what do they do when these sources conflict?

In Lagnado and Sloman's (in preparation) experiment participants had to learn about the connections in a simple computer network. To do so, they sent test messages from a master computer to one of four computers in a network, and then observed which of the other computers also received the messages. They were able to send 100 test messages before being asked about the structure of the network. Participants completed four tasks, each with a different network of computers. They were instructed that there would sometimes be delays in the time taken for the messages to be transmitted from computer to computer. They were also told that the connections, where they existed, only worked 80% of the time. (In fact the probabilistic nature of the connections is essential if the structure of the network is to be learnable from correlational information. With a deterministic network all the connected computers would covary perfectly, so it would be impossible to figure out the relevant conditional independencies.)

Unknown to participants, the networks in each problem had the same underlying structure, and only differed in the temporal order in which the computers displayed their messages. The four different temporal orderings are shown in Figure 2, along with the links endorsed by the participants in the test phase. When the temporal ordering reflected the underlying network structure, the correct model was generally inferred. When the information was presented simultaneously learners did less well (adding incorrect links) but still tended to capture the main links. When the temporal ordering conflicted with the underlying structure, participants erroneously added links that fitted with the temporal order but that did not correspond to the underlying structure.

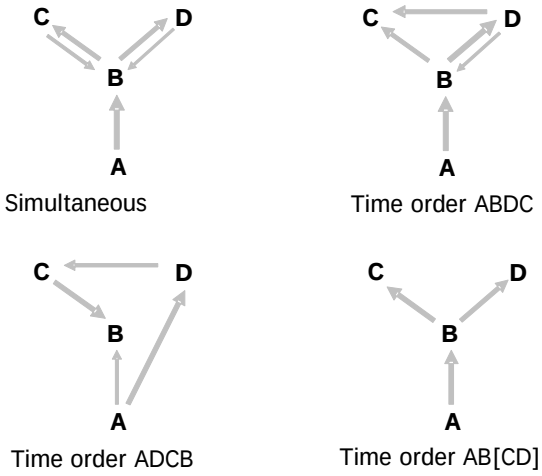

Figure 2. Model choices for the four temporal conditions in Lagnado and Sloman (in preparation). Note: only links endorsed by > 50% participants are shown, and the thickness of the arrows corresponds to the percentage of participants selecting that link (thickest link is = 100%).

In sum, people allowed the temporal ordering to guide their structural inferences, even when this conflicted with the structure implied by the correlational information. However, this did not amount to a total disregard of the correlational information. For example, in the problem with temporal ordering ABDC (top right panel in Figure 2), participants erroneously endorsed the link from D to C (as suggested by the temporal order) but also correctly added the link from A to C. We hypothesize that they first used the temporal ordering to set up an initial model (A→B→D→C). This model would be confirmed by most of the test trials. However, occasionally they saw a test pattern that contradicted this model (A, B, not-D, C). To accommodate this new evidence, they added a link from B to C, but did not remove the redundant link from D to C, because this still fit with the temporal ordering.

Interpreted within the causal-model framework, this study shows that people use both temporal order and correlational cues to infer causal structure. It also suggests that they construct an initial model on the basis of the temporal ordering

(when available), and then revise this model in the light of the covariational information. However, due to the persisting influence of the temporal order cue, these revisions may not be optimal.

Although the reported study highlights how people can be misled by an inappropriate temporal ordering, in many contexts the temporal cue will reliably indicate the correct causal structure. As with other mental heuristics, its fallibility does not undermine its effectiveness in most naturalistic learning situations. It also works best when combined with other cues. In the next section we shall examine how it combines with interventions.

4.3 Intervention

Various philosophers have argued that the core notion of causation involves human intervention (Collingwood, 1940; Hart & Honore, 1983; Von Wright, 1971). It is through our actions and manipulations of the environment around us that we acquire our basic sense of causality. Several important claims stem from this: that causes are potential handles upon the world; that they 'make a difference'; that they involve some kind of force or power. Indeed the language and metaphors of causal talk are rooted in this idea of human intervention on a physical world. More contemporary theories of causality dispense with its anthropomorphic connotations, but maintain the notion of intervention as a central concept (Glymour, 2001; Pearl, 2000; Spirtes et al., 1993; Woodward, 2003).

Intervention is not only central to our notion of causation. It is a fundamental means by which we learn about causal structure. This has been a commonplace insight in scientific method since Bacon (1620) spoke of 'twisting the lion's tail', and was refined into axioms of experimental method by Mill (1843). More recently, it has

been formalized by researchers in AI and philosophy (Spirtes et al., 1993; Pearl, 2000; see Hagmayer et al., this volume).

The importance of intervention in causal learning is slowly beginning to permeate through to empirical psychology. Although it has previously been marked in terms of instrumental or operant conditioning (Mackintosh & Dickinson, 1979), the full implications of its role in causal structure learning had not been noted. This is largely due to the focus on strength estimation rather than structural inference. Once the emphasis is shifted to the question of how people infer causal structure, the notion of an intervention becomes critical.

Informally, an intervention involves imposing a change on a variable in a causal system from outside the system. A strong intervention is one that sets the variable in question to a particular value, and thus overrides the effects of any other causes of that variable. It does this without *directly* changing anything else in the system, although of course other variables in the system can change *indirectly* as a result of changes to the intervened-on variable (a more formal definition is given by Woodward, this volume).

An intervention does not have to be a human action (cf. Mendelian randomization, Davey Smith & Ebrahim, 2003), but freely chosen human actions will often qualify as such. These can range from carefully controlled medical trials to the haphazard actions of a drunk trying to open his front door. Somewhere in between lays the vast majority of everyday interventions. What is important for the purposes of causal learning is that an intervention can act as a quasi-experiment, one that eliminates (or reduces) confounds and helps establish the existence of a causal relation between the intervened-on variable and its effects.

A central benefit of an intervention is that it allows one to distinguish between causal structures that are difficult or impossible to discriminate amongst on the basis of correlational data alone. For example, a high correlation between bacteria and ulcers in the stomach does not tell us whether the ulcers cause the bacteria or vice-versa (or, alternatively, if both share a common cause). However, suppose an intervention is made to eradicate the bacteria (and that this intervention does not promote or inhibit the presence of ulcers by some other means). If the ulcers also disappear, one can infer that the bacteria cause the stomach ulcer and not vice versa.

Intervention aids learning

Can people make use of interventions in order to learn about causal structure? Several studies have compared learning through intervention with learning through observation (Lagnado & Sloman, 2002, 2004; Sobel, 2003; Steyvers et al., 2003). All these studies have shown a distinct advantage for intervention. When participants are able to freely intervene on a causal system they learn its structure better than when they are restricted to passive observation of its autonomous behavior.

What are the factors that drive this advantage? In addition to the special kind of information afforded by intervention, due to the modification of the system under study, interventions can facilitate learning in several other ways. For instance, an intervener has more control over the kind of data they see, and thus can engage in more directed hypothesis testing than an observer. Intervention can also focus attention on the intervened-on variables and its effects. Further, the act of intervention introduces an implicit temporal cue into the learning situation, because interventions typically precede their effects. Interveners may use any of these factors to enhance their learning.

By using yoked designs Lagnado and Sloman (2004, in preparation) ruled out the ability to hypothesis-test as a major contributor in their experiments (though Sobel & Kushnir, 2003, report conflicting results). However, they also showed that the presence of a temporal cue had a substantial effect. When the information about the variables in the causal system was presented in a temporal order that matched the actual causal order (rather than being inconsistent with it) learning was greatly facilitated, irrespective of whether participants were intervening or observing. The authors suggested that in general people might use a temporal order heuristic whereby they assume that any changes that occur subsequent to an action are effects of that action. This can be an effective heuristic, especially if these actions are unconfounded with other potential causes of the observed effects. Such a heuristic can also be used in observation, but is more likely to lead to spurious inferences (because of unavoidable confounding).

An online learning paradigm

Although all of the studies reported so far demonstrate an advantage of intervention, they also reveal low levels of overall performance. Even when learners were able to intervene, many failed to learn the correct model (in most of the experiments less than 40% chose the correct models). We conjecture that this is due to the impoverished nature of the learning environment presented to participants. All of the studies used a trial-based paradigm, in which participants viewed the results of their interventions in a case-by-case fashion. And the causal events under study were represented by symbolic descriptions rather than being directly experienced (cf. Waldmann & Hagmayer, 2001). This is far-removed from a naturalistic learning context. Although it facilitates the presentation of the relevant statistical contingencies, it denies the learner many of the cues that accompany real-world

interventions like spatiotemporal information, immediate feedback, and continuous control.

To address this question, Lagnado and Sloman (in preparation) introduced a learning paradigm that provided some of these cues; participants manipulated on-screen sliders in a real-time environment. Participants had to figure out the causal connections between the sliders by freely changing the settings of each slider, and observing the resultant changes in the other sliders. In these studies the majority of learners (greater than 80%) rapidly learned a range of causal models, including models with four inter-connected variables. This contrasted with the performance of observers, who watched the system of sliders move autonomously, and seldom uncovered the correct model. Thus the benefits of intervention seem to be greatly magnified by the dynamic nature of the task. This reinforces our claim that causal cognition operates best when presented with a confluence of cues and, in particular, that intervention works best when combined with spatiotemporal information.

In addition, in a separate condition many learners were able to make use of double interventions to disambiguate between models indistinguishable through single interventions. For example, when restricted to moving one slider at a time it is impossible to discriminate between a three variable chain A→B→C, and a similar model with an extra link from A to C. However, with an appropriate double intervention (e.g., fixing the value of B, and then seeing whether manipulation of A still leads to a change in C) these models can be discriminated. The fact that many participants were able to do this shows that they can reason using causal representations (cf. Hagmayer et al., this volume). They were able to represent the two possible causal models, and work out what combination of interventions would discriminate between them.

Intervention vs. temporal order

The trial-based experiments by Lagnado and Sloman (2004) show that temporal order plays a substantial role in causal learning. However, the low levels of performance made it difficult to assess the separate influences of intervention and temporal order cues. A subsequent study by Lagnado and Sloman (in preparation) used the slider paradigm to investigate this question. Participants completed six problems, ranging from two-variable to four-variable models. They were divided into three groups: those who could freely intervene on the causal system, those who observed the system's behavior, and those who observed the results of another person's interventions (yoked to the active interveners). Within each group participants were presented with information about the slider values in two temporal orders, either consistent with, or opposite to, the underlying causal structure. The main results are shown in Figure 3 (where the intervention group is denoted by intervention1). There is a clear advantage of intervention (active or yoked) over observation. There is also a clear influence of temporal consistency for the observational and yoked groups, but not for the active interveners. The authors conjectured that the active interveners overcame the inconsistent temporal order cue by (correctly) learning that the variable information was presented in reverse order. To test this they ran a second intervention condition in which the temporally inconsistent time order was randomized rather than reversed (with the constraint that it could never produce a consistent order). The results for this follow-up are also shown in Figure 3 (the new intervention group is intervention2). The interveners now showed a similar decline in performance when information was presented in an inconsistent order. Overall these results confirm that intervention and temporal order provide separate cues to causal structure. They work best, however, in combination,

and this may explain the efficacy of interventions made in naturalistic learning environments.

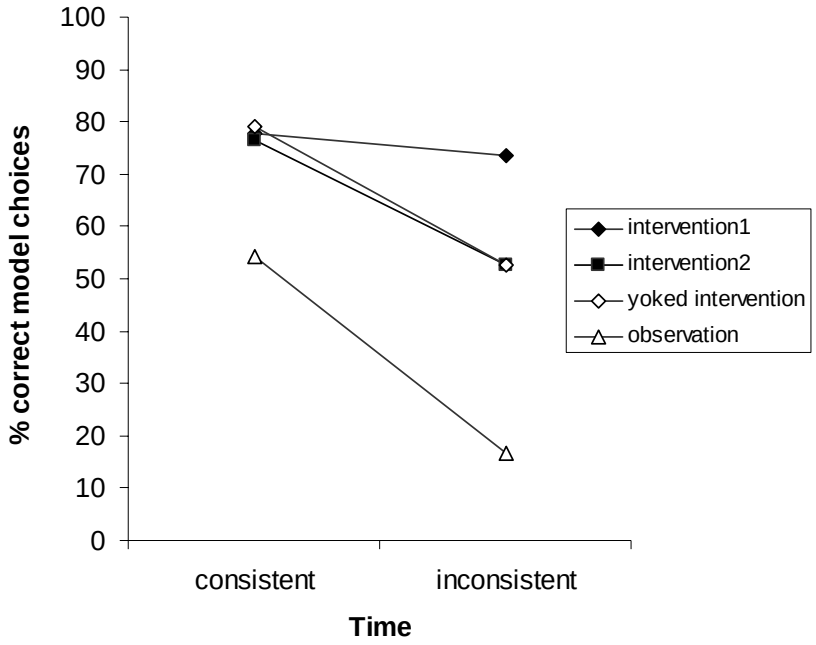

Figure 3. Percent correct model choices in Lagnado and Sloman (in preparation) showing influence of intervention and temporal order. Note: in intervention2 time inconsistent orders were randomized rather than reversed.

4.4 Prior Knowledge

Temporal order is a powerful cue to causality in situations in which we experience causal events on-line. Whenever we directly experience causal events the sequence of the learning input (i.e., learning order) mirrors the asymmetry of causal order (causes generate effects but not vice versa). The correlation between learning order and causal order is so strong in these situations that some theories (e.g., associative learning models) collapse causal order and learning order by assuming that learning generally involves associations between cues and outcomes with cues presented temporally prior to their outcomes (see Shanks & Lopez, 1996; Waldmann, 1996, 2000).

However, whereas nonhuman animals indeed typically experience causes prior to their effects, the correlation between learning order and causal order is often broken when learning is based on symbolized representations of causal events. In fact, most experimental studies of human learning are nowadays carried out on a computer in which cues and outcomes are presented verbally. The flexibility of symbolic representations allows it to present effect information prior to cause information so that learning order no longer necessarily corresponds to causal order. For example, many experiments have studied disease classification in which symptoms (i.e., effects of diseases) are presented as cues prior to information about their causes (i.e., diseases; e.g., Gluck & Bower, 1988; Shanks & Lopez, 1996; Waldmann, 2000, 2001).

Learning order and causal order may also mismatch when the causal events are not readily observable but have to be measured or searched with more complicated procedures. For example, a physician may immediately observe symptoms of a new patient prior to searching for possible causes. Or parents might become aware of school problems of their child prior to finding out about the causes. Thus, although the temporal order of learning events is often a valid cue to causal structure, it is sometimes necessary to override this cue when other cues appear more valid.

Coherence with prior knowledge is a potent cue to causal structure. Regardless of when we observe fever in a patient, our world knowledge tells us that fever is not a cause but rather an effect of an underlying disease. Prior knowledge may be very specific when we have already learned about a causal relation, but prior knowledge can also be abstract and hypothetical. We know that switches can turn on devices even when we do not know about the specific function of a switch in a novel

device. Similarly we know that diseases can cause a wide range of symptoms prior to finding out which symptom is caused by which disease. In contrast, rarely do we consider symptoms as possible causes of a disease.

Prior Knowledge versus Temporal Order

The possible mismatch between causal order and learning order raises the question whether people are capable of ignoring the temporal order of learning events when their prior knowledge suggests a different causal order. Following the framework of causal-model theory, Waldmann and Holyoak (1990, 1992) developed an experimental paradigm addressing this question. In general, learners in different conditions receive identical cues and outcomes in identical learning order. However, based on initial instructions different causal orders are suggested so that in one condition the cues represent causes and the outcomes effects (predictive learning), whereas in the contrasting condition the cues represent effects and the outcomes causes (diagnostic learning).

A recent study by Waldmann (2001) exemplifies this paradigm. In Experiment 2 of Waldmann (2001), participants were told that they are going to learn about new diseases of the blood. In all conditions learners observed learning trials in which they first received information about the presence of a Substance 1 in a patient followed by feedback about the presence of a disease (e.g., Midosis). Other trials showed patients whose blood contained two substances, Substance 2 and 3, which were a sign of a different disease (e.g., Xeritis). Associative learning theories are only sensitive to learning order and would therefore generally predict that the associative strength between Substance 1 and Midosis should be greater than between the two other substances and Xeritis (see Cobos et al., 2002). This so called overshadowing effect falls out of associative learning theories (e.g., Rescorla & Wagner, 1972) which

predict that the two redundant always co-occurring substances compete for predicting Xeritis. Once asymptotic performance is achieved this should lead to either substance contributing only about half of the associative strength needed to correctly predict the disease.

To pit learning order against causal order, Waldmann (2001) created two contrasting conditions: In the *predictive-learning condition* the substances were described as coming from food items, which gives them the status of potential causes of the diseases (see Figure 2). In contrast, in the *diagnostic-learning condition* the substances were characterized as being potentially generated by the diseases, which assigns them the causal status of effects. Although, cues, outcomes, and learning order were identical in both conditions, overshadowing interacted with causal status. Overshadowing was stronger in the predictive than in the diagnostic-learning condition. Similar interactions have also been shown for the related blocking phenomenon (Waldmann & Holyoak, 1992; Waldmann, 2000; Waldmann & Walker, in press).

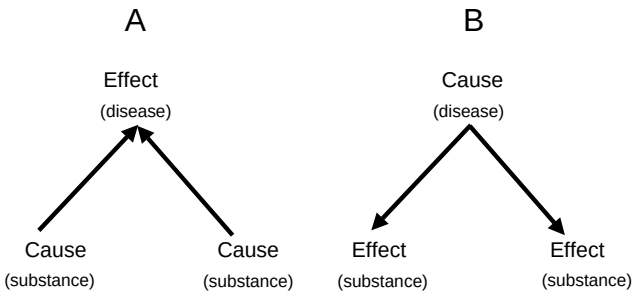

Figure 2. Predictive-learning (A) and diagnostic-learning (B) in Waldmann (2001).

This interaction can be modeled by an account that assumes that people use prior knowledge conveyed to them through the cover stories to form tentative causal models with the structures displayed in Figure 2 (see Waldmann & Martignon, 1998, for a Bayesian causal-model account). These models free learners from learning order as a cue to causality, and allow them to flexibly assign the learning input to the causal variables in the tentative causal model. Thus, in the predictive-learning condition the cues are being mapped to the cause layer and the effects to the outcome layer (Figure 2, A), whereas in the diagnostic-learning condition the cues are being mapped to the effect layer and the outcomes to the cause layer (Figure 2, B). Although prior knowledge generates the structure of the causal models underlying the learning domain, the cover stories made it clear to participants that the causal relations were only hypothetical and needed to be verified by checking the learning data. Thus, in the learning stage the learning input is used to parameterize the model or test whether the hypothesized links are present.

The overshadowing study illustrates this account (Waldmann, 2001). In Experiment 2 learners observed Substance 1 by itself as a deterministic cause (predictive learning) or a deterministic effect (diagnostic learning). However, the situation differed across the two learning conditions for the two redundant substances. In the diagnostic-learning condition the data suggest that each of the two substances is deterministically caused by their common cause, the disease Xeritis (see Figure 2, B). Although there may be alternative unknown diseases also affecting these symptoms, these alternative causes were not mentioned in the instructions so that their potential impact on learners' assessment should be relatively small. By contrast, in the predictive-learning condition, learners were confronted with an ambiguous situation. Here the two substances represented perfectly confounded alternative potential causes

so that it was impossible to determine whether only one of these potential causes was effective, or whether both shared responsibility in generating the common effect, Midosis (see Figure 2, A). Thus, learners should have been uncertain about their causal status, which would lead to a lowering of ratings (i.e., overshadowing). This pattern was indeed found in the study.

Temporal order of learning events was also pitted against causal order in other task domains. In a study on category learning, Waldmann, Holyoak, and Fratianne (1995) have shown that sensitivity to correlations among cues is influenced by the causal status of the cues (see also Rehder & Hastie, 2001; Rehder, 2003a, b). As predicted by Bayesian models, when the cues represent effects within a common-cause model, learners expected cue correlations, whereas statistical independence among cues is expected when they represent multiple causes of a common effect. These expectations influenced how difficult it was for participants to learn about different category structures. Again these findings support the view that learners formed a structural representation of a causal model on the basis of the initial instructions, and then tried to map these models to the learning data (see Waldmann & Martignon, 1998, for the formal details).

Prior Knowledge and Parameter Estimation

Even when causal order and temporal order coincide, temporal order alone is not sufficiently constrained to determine how learning events should be processed. In a stream of learning events, the relevant events need to be parsed first, and then it is necessary to decide how the events are interrelated. Often this problem is solved by assuming that events that are spatiotemporally contiguous (see Section 4.2) are interrelated. But this is not always true. For example, when eating a fish dish we would not view the fish as a cause of a subsequent nausea that occurred within 0.5

seconds of eating the meal. Based on prior knowledge, we expect a longer latency of the causal mechanism. In contrast, we would not relate a light to a button press if there was a latency of 10 seconds between pressing the button and the light going on.

Hagmayer and Waldmann (2002) have shown that prior expectations about temporal delays between causes and effects indeed mediate how causes and effects are interrelated within a stream of events. This selection consequently affects how causal strength is estimated within the data set. Despite observing identical event streams, different assessments of causal strength resulted based on how the stream was parsed and how the events were interrelated prior to assessing the strength of covariation.

Prior assumptions also affect what statistical indicators are chosen to estimate causal strength parameters. When the task is to estimate causal strength between a cause and effect, it is necessary to compute the covariation between these events while holding constant alternative causes that may confound the relation. For example, the strength of the causal influence of smoking on heart disease should ideally be assessed when alternative causes of heart disease (e.g., junk food) are absent or held constant. In contrast, causally irrelevant events, alternative effects of the target cause (within a common-cause model), or events that lie downstream on a causal chain between the target cause and the target effect must not be held constant (Eells, 1991; Pearl, 2000). Otherwise, erroneous parameter estimates might result.

Waldmann and Hagmayer (2001) have shown that participants are indeed sensitive to these normative constraints. In a set of experiments, learners were given identical learning input with three interrelated events. Participants' task was to assess the strength of the causal relation between a given cause and an effect. The causal status of the third event was manipulated by means of initial instructions. The results

showed that learners only tended to hold the third event constant when this event was assumed to be an alternative cause of the target effect. When it was causally irrelevant, an alternative effect of the cause or an intermediate event on a causal chain between cause and effect, participants ignored the status of the third variable. Again this is a case in which temporal order is an insufficient cue because the learning events were presented identically to all participants. The correct parameter estimates depended on prior knowledge about the causal status of the learning events.

Use of Prior Knowledge and Processing Constraints

Processing learning data on the basis of a prior causal model can be demanding. For example, in a diagnostic learning task the learning order of cues and outcomes conflicts with causal order. Also holding constant alternative causes can sometimes be difficult when the presence and absence of the alternative cause alternates so that it is hard to separately store in memory the events in which the confound was present and absent. A number of recent studies have shown that in situations that tax processing capacity, people may incorrectly process the learning data, although in less complex tasks they do better (De Houwer & Beckers, 2003; Tangen & Allan, 2004; Waldmann & Hagmayer, 2001; Waldmann & Walker, in press). Waldmann and Walker (in press) have additionally shown that it is crucial that people have a strong belief in the validity of the causal model; otherwise their learning is dictated by other cues that require less effort to use. These studies show that people have the competence to correctly interrelate causal models and learning data, when they strongly believe in their prior assumptions and when the learning task is within the grasp of their information processing capacity. Otherwise, other cues may dominate.

5. Integrating Fragments of Causal Models

We rarely acquire knowledge about a complex causal network all at once. Rather we learn about these models in a piecemeal fashion. Consider, for example, the search for the causes of ulcer by medical science (see Thagard, 1999, for a detailed description of the history of medical theories of ulcer). It was first thought that ulcers were caused by excessive acid in the stomach, which was caused by stress. Later on scientists found out that excessive acidity is not the cause of many ulcers, but that the majority of ulcers are caused by bacteria (helicobacter pyloris). Additionally, it was discovered by other researchers that some acid-based pain relievers, such as aspirin, might also cause ulcers. As a consequence, an initially simple causal-chain model (stress→ excessive acid →ulcer) was replaced by a more complex causal model (see Figure 1). Theory change occurred as a result of many independent empirical studies that focused on individual links. These individual pieces of knowledge were eventually integrated into a coherent, global causal model that incorporated what we now know about ulcers.

Similarly in everyday life we may independently learn that peanuts cause an allergy, and later discover that strawberries cause the same allergy. Although we may never have eaten peanuts and strawberries together, we could still integrate these two pieces of causal knowledge into a common-effect model. Similarly, we might independently learn about two causal relations in which the same common cause is involved. For example, we may first experience that aspirin relieves headache. Later a physician might tell us that our ulcer is also caused by aspirin. Again, although we may never have consciously experienced the two effects of the common cause together, we can integrate the two fragments into a coherent common-cause model.

What is the advantage of integrating fragments of causal knowledge into a coherent global causal model? Despite representing only the direct causal relations within the model (i.e., causes, effects, and causal arrows), causal models allow us to infer the relation between any pair of events within the model, even when they are not directly causally connected. For example the causal model for aspirin would imply that relief of headache and ulcer should tend to co-occur despite not being causally related to each other. These *structural implications* are a consequence of the patterns of causal directionality inherent in causal models.

Bayes nets provide formal tools to analyze structural implications of causal models (see Pearl, 1988, 2000; Spirtes et al., 1993). The graph of a common-cause model expresses that the two effects are spuriously related (due to their common cause) but become independent, once the state of the common cause is known (see Figure 2, B). This is a consequence of the Markov condition. For example, once we know that aspirin is present, the probability of ulcers is fixed regardless of whether headache is present or absent. Similarly, causal chains imply that the initial cause and the final effect are dependent but become independent when the intermediate event is held constant. Finally, a common-effect model (Figure 2, A) implies independence of the alternative causes, but they become dependent once the common effect is held constant. This is an example of explaining away. Eating peanuts and eating strawberries should normally occur independently. But once we know that someone has an allergy, finding out that they have eaten peanuts makes it less likely that they have also eaten strawberries.

Hagmayer and Waldmann (2000, forthcoming) have investigated the question of whether people are capable of integrating fragments into global causal models in a normative fashion (see also Ahn & Dennis, 2000; Perales, Catena, & Maldonado,

2004). In a typical experiment participants had to learn about the causal relations between the mutation of a fictitious gene and two substances. The two relations were learned on separated trials so that no information about the covariation between the two substances was available. Although the learning input was identical, the underlying causal model differed in different conditions. To manipulate causal models participants were either told that the mutation of the fictitious gene was the common cause of two substances, or they were told that the two substances were different causes of the mutation of the gene. The strength of the causal relations was also manipulated to test whether people are sensitive to the size of the parameters when making predictions.

The main goal of the study was to test under which conditions people are aware of the different structural implications of the common-cause and the common-effect model. A correlation should be expected between the two substances when they were caused by a common cause with the size of the correlation being dependent on the size of causal strength. By contrast, two causes of a common effect should be independent regardless of the size of the causal relations.

To test sensitivity to structural implications, participants were given two tasks: In the first task, participants were given new cases along with information about whether a mutation had occurred or not. Their task was to predict on each trial whether either of the two substances was present or absent. Thus, in the common-cause conditions people predicted the presence or absence of the two effects based on information about the presence or absence of the common cause, in the common-effect condition people diagnosed the presence or absence of either cause based on information about the presence or absence of the common effect. This way, participants made predictions for the two substances they had never observed

together. Across multiple predictions participants generated a correlation between the two substances that could be used as an indicator of sensitivity to the implied correlations. The second task asked participants directly to estimate the conditional frequency of the second substance given that the first substance was present or absent.

The two tasks assess sensitivity to structural implications in different ways. Whereas the second task assessed more *explicit* knowledge of the structural implications of causal knowledge, the first task required participants to use the causal models to predict patterns of events. Thus this task probes sensitivity to structural implications in a more *implicit* fashion. For example, in the common-cause condition a possible strategy would be to run a *mental simulation* of the underlying common-cause model. Whenever the presence of the common cause is stated in the test trial, the two effects could be individually predicted with probabilities that conform to the learned strength of the causal relation. This strategy would yield the normatively implied spurious correlation between the substances although the predictions focused on the individual links between the common cause and either effect. Similarly, in the common-effect condition people could simulate diagnoses of the two causes based on information about the presence or absence of the common effect by running the causal model backward from effect to causes (see Figure 2, A). Simultaneous diagnoses of either cause should make participants aware of the possible competition between the causes. Since either cause suffices to explain the effect, people should be reluctant to predict both causes too often. This would yield correct diagnoses of the patterns of causes without requiring participants to directly reflect on the correlation between alternative causes.

The results of this and other experiments show little sensitivity to the differences between common-cause and common-effect models in the explicit

measure. Although some basic explicit knowledge cannot be ruled out (see also Perales et al., 2004), Hagmayer and Waldmann's (2000, in preparation) experiments show that people do not use the strength parameters to predict the implied correlations very well. By contrast, the implicit tasks revealed patterns that corresponded remarkably well to the expected patterns. Whereas a spurious correlation was predicted in the common-cause condition, the predicted correlation stayed close to zero in the common-effect condition. Hagmayer and Waldmann attribute this competency to mental simulations of causal models.

Further experiments by Hagmayer and Waldmann (in preparation) explored the boundary conditions for these effects. The dissociation between explicit and implicit knowledge disappeared with causal chains in which the individual links were taught separately, and in which the task in the test phase was to predict the final effect based on information about the initial cause (see also Ahn & Dennis, 2000). In this task, both explicit and implicit measures were sensitive to the implied correlation between these two events. This result shows that spurious relations (e.g., between two effects of a common cause) need to be psychologically distinguished from indirect causal relations. Whereas people obviously have little explicit knowledge about spurious relations they may view indirect relations as a subdivided global causal relation. In fact, all direct causal relations can be sub-divided into chains that represent the underlying mechanisms. Thus, combining links of causal chains into a global prediction is easier than deriving prediction for spurious relations.

The implicit task also turned out to be sensitive to boundary conditions. Whereas performance for the common-cause model and the chain model showed fairly robust sensitivity to spurious and indirect relations, it turned out that people's implicit estimates in the common-effect condition are only adequate when the task

required them to predict patterns of causes, as in the experiment described above. In this task the links of the causal models were simulated in parallel, which apparently proved important for making learners aware of the implied competition among the causes. When the task was to first predict the effect based on one cause, and then make inferences about the other cause, people erroneously predicted a spurious correlation between the causes. Probably participants accessed each link consecutively and tended to forget about the possible competition between the causes.

In sum, people are capable of integrating fragments of causal knowledge in a way that corresponds to the normative analyses of Bayes nets. However, this competency is not as robust as the computer models used to implement Bayes nets. It rather depends on a number of task factors that include the type of relation within a causal model, and the specifics of the task.

6. Computational models of learning

Although our main concern has been with how people learn causal structure, the story we have told is linked in important ways to current computational models of inference and learning. For one, the Causal Bayesian network formalism (Spirtes et al., 1993; Pearl, 2000) offers a normative framework for causal representation and inference. And at a qualitative level human inference seems to fit with the broad prescriptions of this theory (see Hagmayer et al., this volume; Sloman & Lagnado, 2004, 2005). The Causal Bayesian network framework also suggests various computational procedures for learning causal structure. These are often grouped into two types – Bayesian methods (Heckerman, Meek & Cooper, 1999) and constraint-based methods (Spirtes et al., 1993). It is instructive to compare and contrast these approaches as models of human learning, in the light of the proposals and empirical evidence surveyed in this chapter.

In short, Bayesian methods assume that learners have some prior belief distribution across all possible causal structures, and update these beliefs as statistical data is gathered. Bayes' rule is used to compute posterior probabilities for each of the possible models given the data, and a best fitting model is derived from this computation. Constraint-based methods work by computing the independencies and dependencies (both conditional and unconditional) in the data set, and then returning the structures that are consistent with these dependencies (for more details see Danks, 2005, forthcoming).

At present these computational models have been used as rational rather than psychological models of human learning (Anderson, 1990; Marr, 1982). They aim to specify what the learner is computing, rather than how they are actually computing it. Both Bayesian methods (Steyvers et al., 2003; Tenenbaum & Griffiths, 2003) and constraint-based methods (Gopnik et al., 2004) have been used for this purpose. A question closer to the concerns of the empirical psychologist, however, is whether these models tell us anything about the psychological or process level of causal learning. What are the mechanisms that people actually use to learn about causal structure?

In their current state these computational approaches seem to both over-estimate and under-estimate the capabilities of human learners. For instance, they over-estimate the computational resources and processing power available to humans in order to make the appropriate Bayesian or constraint-based computations. Bayesian models require priors across all possible models, and Bayesian updating with each new piece of evidence. Constraint-based models require the computation of all the dependencies and independencies in the data, and inference of the set of Markov

equivalent structures. Both methods appear to place insurmountable demands on a human mind that is known to be limited in its processing capacities.

There are potential solutions to these shortcomings. Bayesian methods can be heuristic rather than exhaustive, and constraint-based methods can use more psychologically realistic methods for computing dependencies (Danks, 2005). However, both approaches still need to deal with the basic problem, detailed in this chapter, that there is very little evidence that people who only observe patterns of covariation between events (without further constraints) can induce causal models. In particular, there is no clear evidence that people can use statistical information from triples of events to infer causal models via conditional dependence relations. And this ability seems to lie at the heart of both approaches.

In addition, these computational approaches seem to underestimate human capabilities (or, more precisely, the richness of the environment around them, and their ability to exploit this information). As we have seen throughout this chapter, people make use of various cues aside from statistical data. These cues are typically used to establish an initial causal model, which is then tested against the incoming statistical data. Bayesian approaches have sought to model this by incorporating prior knowledge and assumptions in the learner's prior belief distribution (Tenenbaum & Griffiths, 2003), and thus account for inferences made on very sparse data. But it is not clear how they handle cases where people test just a single model, and then abandon it in favor of an alternative. This kind of discontinuity in someone's beliefs does not emerge naturally from the Bayesian approach[1].

On the face of it constraint-based methods are largely data-driven, so the use of prior knowledge and other assumptions appears problematic. But they too have the

[1] This point was made by David Danks (personal communication).

resources to address this issue. Along with the constraints that stem from the statistical dependencies in the data, they can include constraints imposed by prior knowledge, temporal order information, and other cues. This approach also seems to fit well with the discontinuous and incremental nature of human learning (Danks, 2005, forthcoming).

However, in both cases further work is needed to develop a comprehensive framework that can integrate the diverse constraints and cues to structure (e.g., from temporal ordering, interventions, etc.), and capture the heuristic methods that humans seem to adopt. In particular, this framework needs to be able to combine and trade-off these constraints as new information arrives. For example, although an initial causal model might be based on the assumption that temporal order reflects causal order, a revised model could reject this constraint in the light of statistical data that contradicts it (see Section 4.2).

7. Summary

In this chapter we have argued for several interconnected theses. First, the fundamental way that people represent causal knowledge is qualitative, in terms of causal structure. Second, people use a variety of cues to infer structure aside from statistical data (e.g., temporal order, intervention, coherence with prior knowledge). Third, once a structural model is hypothesized, subsequent statistical data is used to confirm or refute the model, and (possibly) to parameterize it. And the structure of a posited model influences how the statistical data itself is processed. Fourth, people are limited in the number and complexity of causal models that they can hold in mind to test, but they can separately learn and then integrate simple models, and revise models by adding and removing single links. Finally, current computational models of learning need further development before they can be applied to human learning.

What is needed is a heuristic-based model that shares the strengths and weaknesses of a human learner, and can take advantage of the rich causal information that the natural environment provides.

References

Ahn, W., & Dennis, M. (2000). Induction of causal chains. In *Proceedings of the Twenty-Second Annual Conference of the Cognitive Science Society* (pp. 19–24). Mahwah, NJ: Erlbaum.

Anderson, J. R. (1990). *The adaptive character of thought*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Bacon, F. (1620). *Novum Organum*. Chicago: Open Court.

Buehner, M. J., Cheng, P. W., & Clifford, D. (2003). From covariation to causation: A test of the assumption of causal power. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*, 1119-1140.

Buehner, M. J., & May, J. (2002). Knowledge mediates the timeframe of covariation assessment in human causal induction. *Thinking and Reasoning, 8*, 269–295.

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

Cobos, P. L., López, F. J., Cano, A., Almaraz, J., & Shanks, D. R. (2002) Mechanisms of predictive and diagnostic causal induction. *Journal of Experimental Psychology: Animal Behavior Processes, 28,* 331-346.

Collingwood, R. (1940): *An essay on metaphysics*. Oxford: Clarendon Press.

Danks, D., & McKenzie, C. R. M. (under revision). Learning complex causal structures.

Danks, D. (2005). Constraint-Based Human Causal Learning. In *Proceedings of sixth international conference on cognitive modelling*, 342-343. Mahwah, NJ: Erlbaum.

Danks, D. (Forthcoming). Causal learning from observations and manipulations. In M. Lovett & P. Shah (Eds.), *Thinking with data*. Hillsdale, NJ: Erlbaum.

Davey Smith, G., & Ebrahim, S. (2003). 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? *International Journal of Epidemiology, 32*, 1-22.

De Houwer, J., & Beckers, T. (2003). Secondary task difficulty modulates forward blocking in human contingency learning. *Quarterly Journal of Experimental Psychology, 56B*, 345-357.

Eells, E. (1991). *Probabilistic causality*. Cambridge: Cambridge University Press.

Einhorn, H. J., & Hogarth, R. M. (1986). Judging probable cause. *Psychological Bulletin*, *99*, 3-19.

Glymour, C. (2001). *The mind's arrows.* 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*, 1–31.

Gluck, M. A., & Bower, G. H. (1988). From conditioning to category learning: An adaptive network model. *Journal of Experimental Psychology: General*, *117*, 227-247.

Griffiths, T. L., &Tenenbaum, J. B. (forthcoming). Elemental causal induction. *Cognitive Psychology*.

Hagmayer, Y., & Waldmann, M. R. (2000). Simulating causal models: The way to structural sensitivity. In L. Gleitman & A. Joshi (Eds.), *Proceedings of the Twenty-second Annual Conference of the Cognitive Science Society*. Mahwah, NJ: Erlbaum.

Hagmayer, Y., & Waldmann, M. R. (2002). How temporal assumptions influence causal judgments. *Memory & Cognition, 30*, 1128–1137.

Hagmayer, Y., & Waldmann, M. R. (in preparation). Integrating fragments of causal models – Implicit versus explicit sensitivity to structural implications.

Hart, H.L.A., & Honoré, T. (1983). *Causation in the law.* Oxford, Clarendon. 2nd ed.

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

Hume, D. (1748). *An enquiry concerning human understanding*. Oxford, England: Clarendon.

Lagnado, D. A., & Sloman, S. A. (2002). Learning causal structure. *Proceedings of the Twenty-Fourth Annual Conference of the Cognitive Science Society*, (pp.560-565). Mahwah, NJ: Erlbaum.

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

Lagnado, D. A., & Sloman, S. A. (in preparation). Causal order vs. time order.

Mackintosh, N. J., & Dickinson, A. (1979). Instrumental (Type II) conditioning. In A. Dickinson & R. A. Boakes (Eds.), *Mechanisms of learning and motivation* (pp. 143–169). Hillsdale, NJ: Erlbaum.

Marr, D. (1982). *Vision.* San Francisco, CA: W. H. Freeman.

Mill, J. S. (1843/1950). *Philosophy of scientific method*. New York: Hafner.

Neapolitan, R. E. (2003). *Learning Bayesian Networks.* Pearson US.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. San Mateo, CA: Morgan Kaufmann Publishers.

Pearl, J. (2000). *Causality.* Cambridge, England: Cambridge University Press.

Perales, J.C., Catena, A. & Maldonado, A. (2004). Inferring non-observed correlations from causal scenarios: The role of causal knowledge. *Learning and Motivation, 35*, 115-135.

Rehder, B. (2003). A causal-model theory of conceptual representation and categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29*, 1141-1159.

Rehder, B. (2003). Categorization as causal reasoning. *Cognitive Science, 27*, 709-748.

Rehder, B., & Hastie, R. (2001). Causal knowledge and categories: The effects of causal beliefs on categorization, induction, and similarity. *Journal of Experimental Psychology*: General, 130, 323-360.

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.

Shanks, D. R., Pearson, S. M., & Dickinson, A. (1989). Temporal contiguity and the judgment of causality by human subjects. *Quarterly Journal of Experimental Psychology,* 41B, 139–159.

Shanks, D. R. (2004). Judging covariation and causation. In D. Koehler & N. Harvey (Eds.), *Blackwell handbook of judgment and decision making*. Oxford, England: Blackwell.

Shanks, D. R., & López, F. J. (1996). Causal order does not affect cue selection in human associative learning. *Memory and Cognition, 24,* 511-522.

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

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

Sloman, S. A., & Lagnado, D. A. (2005). Do we "do"? *Cognitive Science*.

Sloman, S. A., & Lagnado, D. A. (2004). Causal invariance in reasoning and learning. In B. Ross (Ed.), *The psychology of learning and motivation* (Vol. 44, pp. 287–325). San Diego: Elsevier Science.

Sobel, D. M. (2003). Watch it, do it, or watch it done. Manuscript submitted for publication.

Sobel, D. M., & Kushnir, T. (2003). Interventions do not solely benefit causal learning. *Proceedings of the Twenty-fifth Annual Meeting of the Cognitive Science Society,* (pp. 1100-1105). Mahwah, NJ: Erlbaum.

Spirtes, P., Glymour, C., & Schienes, R. (1993). *Causation, prediction, and search*. New York: Springer-Verlag.

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

Tangen, J. M., & Allan, L. G. (2004). Cue-interaction and judgments of causality: Contributions of causal and associative processes. *Memory & Cognition, 32*, 107-124.

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

Tenenbaum, J. B., & Griffiths, T. L. (2003). Theory-based causal inference. *Advances in Neural Information Processing Systems 15* (pp. 35-42). Cambridge, MA: The MIT Press.

Thagard, P. (1999). *How scientists explain disease*. Princeton, NJ: Princeton University Press.

von Wright, G. H. (1971). *Explanation and understanding.* Ithaca, NY: Cornell University Press.

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

Waldmann, M. R. (1996). Knowledge-based causal induction. In D. R. Shanks, K. J. Holyoak, & D. L. Medin (Eds.), *The psychology of learning and motivation* (Vol. 34, pp. 47–88). San Diego, CA: Academic Press.

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

Waldmann, M. R., & Hagmayer, Y. (2001). Estimating causal strength: The role of structural knowledge and processing effort. *Cognition, 82*, 27–58.

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

Waldmann, M. R., & Holyoak, K. J. (1990). Can causal induction be reduced to associative learning? In *Proceedings of the 12th Annual Conference of the Cognitive Science Society* (pp. 190-197). Hillsdale, NJ: Erlbaum.

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

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

Waldmann, M. R., & Walker, J. M. (in press). Competence and performance in causal learning. *Learning & Behavior.*

Waldmann, M. R., Holyoak, K. J., & Fratianne, A. (1995). Causal models and the acquisition of category structure. *Journal of Experimental Psychology: General*, *124*, 181-206.

Wasserman, E. A., Chatlosh, D. L., & Neunaber, D. J. (1983). Perception of causal relations in humans: Factors affecting judgments of response-outcome contingencies under free-operant procedures. *Learning and motivation*, *14*, 406-432.

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