Mackintosh, N. J. (1975). A theory of attention: Variations in the associability of stimuli with reinforcement. *Psychological Review*, **82**, 532–552.

Marr, D. (1982). *Vision: A computational approach*. San Francisco, CA: Freeman and Co.

Medin, D. L., & Bettger, J. G. (1991). Sensitivity to changes in base-rate information. *American Journal of Psychology*, **40**, 175–188.

Medin, D. L., & Edelson, S. M. (1988). Problem structure and the use of base-rate information from experience. *Journal of Experimental Psychology: General*, **117**, 68–85.

Miller, R. R., & Matute, H. (1996). Biological significance in forward and backward blocking: Resolution of a discrepancy between animal conditioning and human causal judgment. *Journal of Experimental Psychology: General*, **125**, 370–386.

Minka, T. (2001). *A family of algorithms for approximate Bayesian inference*. PhD thesis, Massachusetts Institute of Technology.

Neal, R. M., & Hinton, G. E. (1998). A view of the EM algorithm that justifies incremental, sparse and other variants. In M. I. Jordan (Ed.), *Learning in graphical models* (pp. 355–368). Kluwer Academic Publishers.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: The effectiveness of reinforcement and non-reinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical conditioning, 2: Current research and theory* (pp. 64–69). New York: Appleton Century-Crofts.

Shanks, D. R. (1985). Forward and backward blocking in human contingency judgement. *Quarterly Journal of Experimental Psychology*, **37B**, 1–21.

Simon, H. (1957). A behavioral model of rational choice. In *Models of man, social and rational: Mathematical essays on rational human behavior in a social setting*. Wiley, New York.

Treebushny, D., & Madsen, H. (2005). On the construction of a reduced rank square-root Kalman filter for efficient uncertainty propagation. *Future Generation Computer Systems*, **21**, 1047–1055.

Waldmann, M. R., Cheng, P. W., Hagmayer, Y., & Blaisdell, A. P. (2007). Causal learning in rats and humans: A minimal rational model (this volume).

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

Waterhouse, S., MacKay, D., & Robinson, T. (1996). Bayesian methods for mixtures of experts. *Advances in Neural Information Processing Systems*, **8**, 351–357.

Yu, A. J., & Dayan, P. (2003). Expected and unexpected uncertainty: ACh and NE in the neocortex. In *Advances in Neural Information Processing Systems* (Vol. 15). Cambridge, MA: MIT Press.

Yu, A. J., & Dayan, P. (2005). Uncertainty, neuromodulation, and attention. *Neuron*, **46**, 681–692.

Chapter 20

Causal learning in rats and humans: a minimal rational model

Michael R. Waldmann,
University of Göttingen, Göttingen, Germany
Patricia W. Cheng,
University of California, Los Angeles, USA
York Hagmayer, and
University of Göttingen, Göttingen, Germany
Aaron P. Blaisdell
University of California, Los Angeles, CA, USA

Introduction

People's ability to predict future events, to explain past events, and to choose appropriate actions to achieve goals belongs to the most central cognitive competencies. How is knowledge about regularities in the world learned, stored, and accessed? An intuitively plausible theory that has been developed in philosophy for many centuries assumes that *causality* is the 'cement of the universe' (Mackie, 1974), which underlies the orderly relations between observable events. According to this view some event types, causes, have the capacity or power to generate their effects. To be a successful agent we need to have causal representations that mirror the causal texture of the world.

The philosopher David Hume questioned this view in his seminal writings (Hume, 1748/1977). He analyzed situations in which we learn about causal relations, and did not detect any empirical input that might correspond to evidence for causal powers. What he found instead was repeated sequence of a pair of spatio-temporally contiguous events, but nothing beyond. Therefore he concluded that causality is a cognitive illusion triggered by associations. Hume did not question that we believe in causal powers, he merely argued that there is nothing in our experiential input that directly corresponds to causal powers.

The psychology of learning has adopted Hume's view by focusing on his analysis of the experiential input. According to many learning theories, causal predictions are driven by associative relations that have been learned on the basis of observed covariations between events (e.g., Allan, 1993; Shanks & Dickinson, 1987). Similar to Pavlov's dog, which has learned to predict food when it hears a tone (i.e., classical conditioning), or to a rat's learning that a lever press produces food (i.e., instrumental conditioning), we learn about predictive relations in our world. There is no need for the concept of causality in this view. Thus, following the epistemology of logical positivism the concept of causality was dropped altogether and replaced by predictive relations exhibited in covariational patterns between observable events.

What do we gain by having causal representations beyond what we already can do with predictive relations gleaned from learning data? Developing earlier work on causal inference (e.g., Goodman, 1983; Kant, 1781/1965; Skyrms, 2000), philosophers and psychologists have analyzed this question in great depth in the past decades and have pointed to several crucial differences (see Pearl, 1988, 2000; Spirtes *et al.*, 1993; Woodward, 2003): (1) If we had no causal knowledge we could not represent the difference between causal and spurious statistical relations, such as the relation between barometers and the weather. Barometers covary with the weather as does smoking with heart disease. However, the first relation is spurious due to a common cause, atmospheric pressure, whereas the second describes a direct causal relation. Hence, if we mechanically change the reading of the barometer, the weather will not be affected, whereas giving up smoking will decrease the likelihood of heart disease. This distinction is crucial for planning actions (see Woodward, 2003). We can generate events by intervening in their causes, whereas interventions in spurious correlates are ineffective. (2) Another important aspect of causality is its inherent directionality. Causes generate effects but not vice versa. For example, the thrusting position of a fist on a pillow causes the indentation in the pillow, rather than vice versa. In contrast, covariations are undirected and therefore do not allow us to make informed inferences about the outcomes of interventions. (3) A final example of the advantages of causal representations is their parsimony when multiple events are involved. For example, learning predictive relations between six events requires us to encode 15 pairwise covariations. Only some of the necessary information may have been made available to learners, however. In contrast, causal models allow us to form more parsimonious representations and make informed guesses about covariations we may never have observed. For example, if we know that one event is the common cause of the other five, we can infer all 15 covariations from knowledge of the causal strength between the cause and each of its five effects (see Pearl, 1988, 2000; Spirtes *et al.*, 1993).

Following Hume, learning theory has focused on the covariations inherent in the learning input, and has neglected how covariations give rise to causal representations. The basic claim was that knowledge about causal relations is nothing more than knowledge of covariations. However, there is another route that can be traced back to Kant's (1781/1965) view of causality. Hume, who did not deny the possibility of hidden causal powers, was indeed right when he pointed to covariations as the primary experiential input suggesting the existence of causal relations. However, his empiricist epistemology was mistaken. As many philosophers of science have revealed, apart from concepts referring to observable events, our theories also contain theoretical

concepts that are only indirectly tied to the observable data (see Glymour, 1980; Quine, 1960; Sneed, 1971). Thus, it is possible to grant that we only have covariational data to support causal hypotheses, while retaining the view that we go beyond the information given and use covariations along with background assumptions to induce genuinely *causal* relations.

Cheng (1997) was the first to take this path in psychology. She has developed a theory (power PC theory), which formalizes how we can infer unobservable causal powers from covariations (see also Buehner & Cheng, 2005). According to this view, we enter the learning process with abstract assumptions about causes generating or preventing effects in the potential presence of hidden causal events. These assumptions combined with learning input allow learners to induce the causal power of events.

Causal-model theory (Waldmann & Holyoak, 1992) whose focus is on more complex causal models similarly has stated that people interpret covariations in light of prior assumptions about causal structures. A consequence of this view, supported in numerous empirical studies, is that identical learning input may lead to different causal representations depending on the characteristics of prior assumptions (see Waldmann *et al.*, 2006, for an overview).

Most recently, causal Bayes net theory has been proposed as a psychological theory of causal cognitions (Gopnik *et al.*, 2004; Sloman, 2005). Whereas power PC and causal-model theory were developed as psychological theories, causal Bayes net theory was originally developed by philosophers, computer scientists, and statisticians as a rational tool for causal discovery in empirical sciences (see Pearl, 1988; Spirtes *et al.*, 1993). Thus, primarily this approach aimed at developing a complex, normative theory of causal induction, and only secondarily claimed to be a psychological theory of everyday learning.

Given that the majority of learning theories have asserted that causal learning can be reduced to forming associations, one of the main goals in the empirical studies of power PC and causal-model theory was to test these theories against the predictions of associative theories (see Cheng, 1997; Buehner & Cheng, 2005; Waldmann, 1996; Waldmann *et al.*, 2006, for overviews). For example, Buehner *et al.* (2003) tested a novel pattern of an influence of the base rate of the effect on judgments of causal strength predicted by the power PC theory and no other theories. They showed that when a question measuring estimated causal strength is unambiguous, the results supported the key prediction of power PC theory that people use estimates of causal power to assess causal strength (see also Wu & Cheng, 1999; Liljeholm & Cheng, 2007). Waldmann and colleagues have shown that people are sensitive to causal directionality in learning (e.g., Waldmann, 2000, 2001) and to the difference between causal and spurious relations (Waldmann & Hagmayer, 2005). All these findings are inconsistent with the predictions of associative learning theories.

With the advent of causal Bayes net theory there is a major new competitor for power PC and causal-model theory, which often makes similar predictions as these theories. In fact, some have argued that previously developed theories such as power PC theory can be modeled as a special case of causal Bayes nets (see Glymour, 2001). We therefore think it is time to take a closer look at discriminating between different computational theories of causal reasoning.

Developing and testing rational models: the dominant view

Thus far, all theories of causal cognitions are developed at the computational level, which, according to Marr's (1982) famous distinction, is concerned with the goals and constraints rather than the algorithms of computations. Moreover, all theories share the view that a rational analysis of what an organism should compute should be the starting point of a successful theory in this field.

Anderson's (1990) book on rational models has been one of the main influences of the current collection. In the first chapters of this book he proposed a methodological strategy for developing rational models, which will provide the starting point of our discussion (see also Chater & Oaksford, 2004). Anderson (1990) motivates rational modeling by pointing to the problems of empirically identifying theories at the implementation level. In psychology we use observable inputs and outputs to induce unobservable mechanisms. Theoretically all mechanism hypotheses are equivalent that generate the same input-output function. This, according to Anderson (1990), leaves us with the problem of the unidentifiability of psychological theories at the mechanism level. Whenever such theories compute identical input–output functions a decision between them is impossible.

An alternative strategy, according to Anderson (1990), is to abandon the search for mechanisms and focus on rational modeling. He postulates six steps in developing a rational model: (1) We need to analyze the goals of the cognitive system, and (2) develop a formal model of the environment to which the system is adapted. (3) Psychology only enters in the form of *minimal* assumptions about computational limitations. These assumptions should be minimal, according to Anderson, to guarantee that the analysis is powerful in the sense that the predictions mainly flow from an analysis of the goals of the cognitive system and the environment and do not depend on assumptions about (unidentifiable) mechanisms. Then (4) a rational model is developed that derives the optimal behavioral function given the stated constraints of the environment and the organism. (5) Finally, these predictions can be compared with the results of empirical behavioral studies. (6) If the predictions are off, the process iterates by going back and revising previous analyses.

This view has been very popular in causal reasoning research, especially in the Bayesian camp. For example, Steyvers *et al.* (2003) defend the priority of rational analysis over theories of psychological implementation. They argue that their model attempts to explain people's behavior 'in terms of approximations to rational statistical inference, but this account does not require that people actually carry out these computations in their conscious thinking, or even in some unconscious but explicit format.' (p. 485). After this statement they acknowledge that simple heuristics might also account for their findings. Similarly, Gopnik *et al.* (2004) pursued the goal to show that the inferences of preschoolers were consistent with the normative predictions of causal Bayes nets, while ignoring that simpler and less powerful causal approaches can account for many of the presented findings.

Thus, there is a tendency of some researchers in this field to focus on the global fit between a single rational model and observed behavior. Alternative theories are either neglected or reinterpreted as possible implementations of the rational account. In our view, it is time to reconsider the relation between rational models and empirical evidence, and revisit the research strategy Anderson (1990) has proposed.

The indeterminacy of rational models

The underdetermination of psychological theories by the data has been one of the driving forces behind Anderson's (1990) rational analysis approach. However, in our view this argument is not restricted to theories at the mechanism level. The underdetermination problem is a general issue for empirical sciences regardless of whether they study the mind or environmental processes (Quine, 1960). In most areas, multiple theories compete, and it is far from clear whether a unique theory will emerge as a winner. Let us revisit some of Anderson's methodological steps in light of this problem in the area of causal reasoning, and show that there is theory competition at each step:

Step 1 requires an analysis of the goals of the cognitive system. It can easily be seen that in causal reasoning research the goal specifications have been highly dependent on the theory that is endorsed by the researcher. An associationist will see the ability to predict events as the primary goal of cognitive systems; somebody who sees causal forces and mechanisms as the basis of causality will instead choose the understanding of causal systems as primary; finally, a causal Bayes net researcher might focus on the goal of representing interventions and observations within a unified causal representation. Of course, all these approaches might be partially correct. But these examples show that there is theory dependence already at the level of the postulation of goals.

Step 2 focuses on the analysis of the environment. Again research on causality provides an excellent example for the theory-ladenness of environmental theories. Causal Bayes net theory is a recent example of a theory whose primary goal was to provide a framework for describing and discovering causal relations in the environment (Pearl, 2000; Spirtes *et al.*, 1993). However, apart from causal Bayes nets there is a wealth of alternative theories of causality which are in part inconsistent with each other but still claim to provide a proper representation of causal relations in the world (see Cartwright, 1989, 2004; Dowe, 2000; Shafer, 1996).

Step 4, the development of a rational model, is clearly dependent on the model of the environment, and is therefore subject to the same constraints. Causal Bayes net theory is a good example of this dependence as it has simultaneously been proposed as a psychological theory (Gopnik *et al.*, 2004) and as a theory of scientific discovery of causal models in the environment (Spirtes *et al.*, 1993). However, other psychological theories of causality were similarly influenced by normative models. Associative accounts such as the probabilistic contrast model (Cheng & Novick, 1992) have predecessors in philosophy (Suppes, 1970; Salmon, 1980) as have psychological theories (Ahn *et al.*, 1995; Shultz, 1982) focusing on causal mechanisms (Dowe, 2000; Salmon, 1984).

The main goal of the present section is to show that it is premature to expect that a careful analysis of the goals of the cognitive system and the environment will generate a unique rational model. The recent debate on the proper rational model for logical or probabilistic reasoning is a good example of how different assumptions may lead to competing theories (see Oaksford & Chater, 2007). We have argued that the steps postulated by Anderson (1990) are tightly constrained by each other. Goals, environment and cognitive systems need to be modeled as a whole in which all components

influence each other, and jointly should be confronted with empirical data. A rational model for the aplysia will surely look different from one for humans. Consequently, we have to be concerned with the possibility of multiple competing rational models that need to be tested and evaluated.

Minimal rational models as a methodological heuristic

We will defend the position that it is useful to consider whether there are alternative rational theories which are less computationally demanding while still fully accounting for the data (see also Daw *et al.*, this volume, for a different but similarly motivated approach). This methodological heuristic we will call *minimality* requirement. Given the indeterminacy at all levels, it is clear that rational models, just like models at other levels, need to be empirically tested. Due to the potential tradeoffs between goals, environment, (innate and acquired) learning biases and information processing limitations, different rational models can be developed and will therefore compete. How can competing rational models be tested? We will discuss some general principles:

(1) The more psychological evidence we consider, the higher the likelihood that we will be able to empirically distinguish between theories. For example, causal Bayes net theory (Gopnik *et al.*, 2004) requires sensitivity to conditional dependence and independence information, whereas alternative theories do not. Showing that people can or cannot pick up conditional dependency information might therefore be relevant for distinguishing between theories. Of course, answering the question of what computations organisms can accomplish is not always easy. For example, many psychologists believed that we cannot, explicitly or implicitly, compute multiple regression weights until a theory, the Rescorla–Wagner model (1972), was developed which shows how such weights can be computed with fairly easy computational routines.

(2) A minimal model allows us to understand better which conclusions are warranted by the evidence and which not. Moreover, they give us a better understanding of what aspects of theories are actually empirically supported, and which are in need of further research. Minimality is a particularly useful heuristic when theories that are hierarchically related compete with each other. For example, power PC theory can be modeled as a special case of causal Bayes nets (see Glymour, 2001). However, this does not mean that all the evidence for power PC theory immediately is inherited by causal Bayes net theory, because this more complex theory may exaggerate the computational capacities of organisms.

(3) Empirical tests of rational models proposed in the literature often blur the distinction between rational models of scientific discovery and of a rational model of the mind. Causal Bayes net theory is an extreme example, as virtually the same model has been postulated for both areas. However, due to different information processing constraints of computers versus humans, a rational model in Artificial Intelligence will certainly be different from one developed in psychology. The distinction between the normative and the psychological is particularly important when it comes to the question of how heuristics or psychological theories relate to rational models. Often it is argued that rational models let us understand what heuristics try to compute. This is certainly useful as long as it is clear that the rational model merely provides a normative analysis of the situation to which an organism adapts rather than a computational model of the mind. For example, the Bayesian inversion formula can be seen as a tool to compute normative responses in a diagnostic judgment task. But that does not mean that the availability heuristic (Tversky & Kahnman, 1973) should be regarded as an implementation of the normative formula. Heuristics and rational models may lead to similar judgments in a wide range of cases; nevertheless they compute different functions. The goal of minimal rational modeling is to discover the function people are actually computing (see also Danks, this volume). Then it might be informative to compare the predictions of minimal rational models with normative rational models.

Causal learning as a test case

Research on causal learning represents an ideal test case for the question of how rational models should be evaluated. In the past decade several theories have been proposed that compete as rational accounts of causal learning. Since it is not possible to discuss all theories, we will focus on three approaches and discuss them on the basis of recent evidence from our laboratories:

(1) *Associative Theories.* Standard associative accounts of causal learning (e.g., Rescorla-Wagner, 1972) will serve as a base-line for our discussion. To demonstrate that human or nonhuman animals are indeed using *causal* representations, it is necessary to show that the obtained experimental effects cannot be explained with a simpler associative account such as merely predictive learning. According to associative theories events are represented as *cues* and *outcomes* rather than causes and effects (see also Waldmann, 1996; Waldmann *et al.*, 2006). Cues are events that serve as triggers for outcome representations regardless of whether they represent causes or effects. Thus, associative theories are insensitive to causal directionality. Moreover, it is assumed that learning is sensitive to observational covariations rather than causal power (see Cheng, 1997). Thus, there is no distinction between covariations based on spurious (e.g., barometer-weather) as opposed to causal relations (atmospheric pressure-barometer). Finally, covariation knowledge may be acquired between different observable events (i.e., classical conditioning) or between acts and outcomes (i.e., instrumental learning).

(2) *Causal Bayes Nets.* Currently different variants of causal Bayes nets are being developed, which compete with each other (see Gopnik & Schulz, 2007, for an overview). We are going to focus on the version proposed by Gopnik *et al.* (2004). In this framework causal models are represented as directed acyclic graphs, which contain nodes connected by causal arrows. 'Acyclic' means that the graph does not contain loops. It is assumed that the graphs satisfy the Markov condition which states that for any variable X in a set of variables S not containing direct or indirect effects of X, X is jointly independent of all variables in S conditional on any set of values of the set of variables that are direct causes of X. An effect of X is a variable that is connected with a single arrow or a path of arrows pointing from X to it.

Figure 20.1 shows an example of three basic causal models. Model A (left) represents a common-cause model in which a common cause (e.g., atmospheric pressure) both causes effect_1 (e.g., barometer) and effect_2 (e.g., weather). The two direct causal links imply covariations between the common cause and either effect. Moreover, the Markov condition implies that the two effects should be spuriously correlated but become independent conditional on the states of their common cause. Model B (middle) represents a causal chain, which has similar implications. The initial cause should covary with effect_1 and effect_2, which is caused by effect_1. Due to the Markov condition, the cause and effect_2 should become independent conditional on effect_1. Finally, Model C (left) represents a common-effect model. In the absence of further external common causes of the two causes 1 and 2, these causes should covary with their joint effect but be mutually marginally independent. However, the causes should become dependent conditional on their common effect.

An important claim of Gopnik *et al.* (2004) is that people should be capable of inducing causal structure from conditional dependence and independence information. Again the Markov assumption along with additional assumptions (e.g., faithfulness) is central for this achievement. Gopnik *et al.* (2004) discuss two Bayesian induction strategies. According to *constraint-based learning* people should analyze triples of events (such as in Fig. 20.1) within causal models and select between causal models on the basis of conditional dependence and independence information. Sometimes this will yield several (Markov equivalent) alternatives. Additional cues (e.g., temporal order information) may help to further restrict the set of possibilities. An alternative to this bottom-up approach are *Bayesian algorithms*, which assign prior probabilities to possible causal models, which are updated by the application of Bayes' theorem given the actual data. Both methods rely on conditional dependence and independence information implied by the Markov condition.

Apart from allowing us to predict events, causal models can also be used to plan actions. Unlike associative theories, causal models are capable of representing the relation between inferences based on observations of events and inferences based on interventions in these events (see also Hagmayer *et al.*, 2007). For example, the common-cause model depicted in Fig. 20.1A implies that the observation of effect_1 (e.g., barometer) allows for inferring the state of effect_2 (weather) based

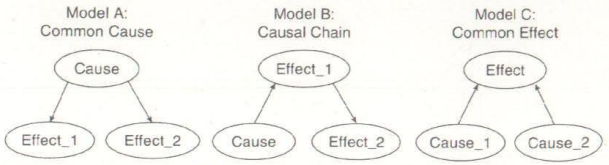

Fig. 20.1. Three types of acyclic causal models connecting three events.

on the diagnostic link between effect_1 (atmospheric pressure) and its cause, and the predictive link between the cause and effect_2. However, manipulating the reading of the barometer by tampering with it should not affect effect_2. Causal Bayes nets allow for modeling this difference by modifying the structure of the graph (Pearl, 2000; Spirtes *et al.*, 1993; Woodward, 2003). Deterministic manipulations of effect nodes render the state of these nodes independent of its causes, as long as some plausible boundary conditions apply (e.g., independence of the instrumental action with the relevant events of the causal models). This can be modeled by removing the arrow between the cause and the manipulated event, which Pearl (2000) vividly called graph surgery (see Fig. 20.2).

The possibility of representing both observational and interventional inferences within a single causal model is one key feature of causal Bayes nets that render them causal. The fact that interventions often imply modifications of causal models turns interventions into an additional powerful tool to induce causal structure. Learning can capitalize from both observational and interventional information and combine these two components during learning (see Gopnik *et al.*, 2004).

(3) *Single-effect Learning Model.* Given our interest in minimal rational models it is useful to test the simplest possible theory of causal learning as an alternative account. Buehner and Cheng (2005) have proposed that organisms primarily focus on evaluating single causal relations during learning. The individual links are integrated into a causal model or causal map (Gopnik *et al.*, 2004; Waldmann & Holyoak, 1992). Should several causal relations contain overlapping events it is possible to make inferences across complex causal networks by chaining the links. The focus on evaluating a single causal relation does not imply that causes of the same effect *e* that are not currently evaluated are ignored; accounting for *e* due to causes other than the candidate *c* is an essential part of inferences about the

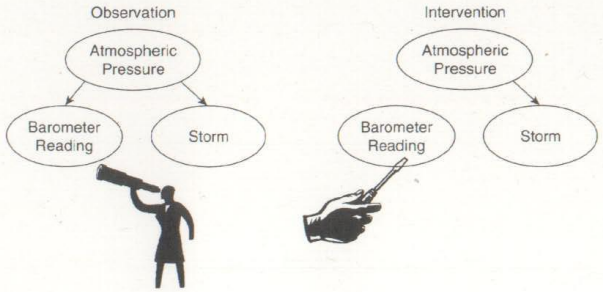

Fig. 20.2. Observing an effect (left) versus intervening in an effect (right) of a common cause: While an observation of an effect allows inferring the presence of its cause, an intervention in the same variable renders this variable independent of its cause. See text for details.

relation between *c* and *e*. Thus, the common-effect structure (e.g., Fig. 20.1C) is the basic unit in which learning occurs, as has been assumed by previous psychological learning theories (e.g., Cheng, 1997; Griffiths & Tenenbaum, 2005; Rescorla & Wagner, 1972). A single-effect learning strategy may be an effective default strategy because, in contrast to the typical wealth of data mined by Bayes nets algorithms, information available to humans and other species regarding relevant causal variables may often be very limited. The information available may be further impoverished by the reasoner's memory and attention constraints. The present chapter will review and explain how, under causal-power assumptions (Cheng, 1997), (1) complex causal models (Waldmann & Holyoak, 1992) are constructed via single-effect learning, and (2) making predictive and diagnostic link-by-link inferences based on the models account for observational and interventional inferences within complex causal networks.

Review of causal model construction via single-effect causal learning

Cheng's (1997) power PC theory provides an account of the learning of the strength of the primary unit of causality—a causal relation between a single candidate cause and a single effect. Whereas associative theories merely encode observable covariations, causal relations do not primarily refer to observable statistics but unobservable theoretical entities. To estimate the causal strength of these unobserved causal relations several assumptions need to be made. The power PC theory partitions all causes of effect *e* into the candidate cause in question, *c*, and *a*, a composite of all observed and unobserved causes of *e*. The unobservable probability with which *c produces e* is termed *generative* power, represented by q_c. The generative power of the composite *a* is analogously labeled q_a. On the condition that *c* and *a* influence *e* independently, it follows that

$$P(e|c) = q_c + P(a|c) \cdot q_a - q_c \cdot P(a|c) \cdot q_a \quad (1), \text{ and}$$

$$P(e|\sim c) = P(a|\sim c) \cdot q_a \quad (2).$$

Equation (1) implies that effect *e* is either caused by *c*, by the composite *a*, or by both, assuming that *c* and *a* produce *e* independently. The difference between $P(e|c)$ and $P(e|\sim c)$ is called ΔP, which is a frequently used measure of covariation in learning research. Thus, from Equation 2, it follows that

$$\Delta P = q_c + P(a|c) \cdot q_a - q_c \cdot P(a|c) \cdot q_a - P(a|\sim c) \cdot q_a \quad (3).$$

Equation (3) shows why covariations do not directly reflect causality. There are four unknowns in the equation. The lack of a unique solution for q_c, the desired unknown, corresponds to the intuitive uncertainty regarding q_c in this situation: if we observe the presence of a candidate cause *c* and its effect *e*, we do not know whether *e* was actually caused by *c*, by *a*, or by both. If *c* and *a* are perfectly correlated, we may observe a perfect covariation between *c* and *e*, and yet *c* may not be a cause of *e* because the confounding variable *a* may be the actual cause. The learner therefore

restricts causal inference to situations in which *c* and *a* occur independently; that is, there is *no confounding*. In that special case, (3) reduces to (4):

$$q_c = \Delta P / (1 - P(e|\sim c)) \quad (4)$$

The above analysis holds for situations in which $\Delta P \geq 0$. A similar derivation can be made for situations in which $\Delta P \leq 0$, and one evaluates the *preventive* causal power of *c*.

The 'no confounding' prerequisite that follows from Equations 3 and 4 explains why interventions have special status as a method for inducing causal power. Interventions typically are assumed to occur independently of the other causes of the target event. This prerequisite also explains why when an intervention is believed to be confounded (e.g., placebo effect; see Buehner & Cheng, 2005; Cheng, 1997), it is not different from any other confounded observation; in this case interventions do not have any special status. Note that to satisfy the 'no confounding' assumption, one does not need to know the identities of other causes or observe their states; one only needs to know that these causes, whatever they may be, occur independently of the candidate cause (e.g., consider the case of random assignment to a treatment and a no-treatment group). Confounding may of course be due to observed alternative causes as well. Research on confounding by observed causes has shown that people are aware of the confounding and therefore tend to create independence by holding the alternative cause constant, preferably in its absent value (see, for example, Waldmann & Hagmayer, 2001; Spellman, 1996).

Thus, while the focus is on the learning of a single causal relation, the possibility of the effect due to other causes is acknowledged. (Information about multiple candidate variables is of course required when one evaluates a conjunctive candidate cause, one that involves a combination of variables.) The single-effect learning theory explains why causal learning can proceed even when one has explicit knowledge of the states of only two variables, the candidate cause and the target effect. The work on causal Bayes nets, in which 'no confounding' is not a general prerequisite for causal learning, have not considered the role of this prerequisite in the case of link-by-link single-effect learning, arguably the most common type of biological learning.

In summary, the basic unit of causal analysis is a common-effect network with an observable candidate cause, alternative hidden or observable causes, and a single effect. These three event types allow organisms to go beyond covariational information and estimate theoretical causal entities.

Causal directionality

A key feature of causal relations is their inherent causal directionality. Causes generate effects but not vice versa. Correct assessments of causal power require the distinction between causes (*c, a*) and effect (*e*). However, it is a well-known fact that causal directionality cannot be recovered from covariation information between two events alone. For example, a flagpole standing on a beach covaries with its shadow on the sand as does the shadow with the flagpole, but the flagpole causes the shadow rather than vice versa. Covariations are symmetric whereas causal power is directed.

Learners' sensitivity to causal directionality has been one of the main research areas of causal-model theory (see Waldmann, 1996; Waldmann *et al.*, 2006, for overviews). According to this theory people use non-statistical *cues* to infer causal directionality (see Lagnado *et al.*, 2007). These cues, although fallible, provide the basis for hypotheses regarding the distinction between causes and effects. What cues are typically used?

Interventions are arguably the best cue to causal directionality. Manipulating a variable turns it into a potential cause, and the change of subsequent events into potential effects. Interventions are particularly useful if they are not confounded (i.e., independent), which may not always be the case, as mentioned earlier. Interventions, particularly unconfounded ones, are not always available. *Temporal order* is another potent cue. Typically cause information temporally precedes effect information. However, the phenomenal representational capacities of humans allow for a decoupling between temporal and causal order. For example, a physician may see information about symptoms prior to the results of tests reflecting their causes, but still form a correct causal model. Research on causal-model theory has shown that humans are indeed capable of focusing on causal order and disregarding temporal cues in such situations (e.g., Waldmann & Holyoak, 1992; Waldmann *et al.*, 1995; Waldmann, 2000, 2001). *Coherence* with prior knowledge is a further potent cue to causal directionality (see also Lien & Cheng, 2000). For example, we know that electrical switches are typical causes even when we do not know what a particular switch causes in a learning situation. Prior knowledge may finally be invoked through *communication*. Instructions may teach us about causal hypotheses.

Diagnostic causal inference under causal-power assumptions

Cheng (1997) and Novick and Cheng (2004) focused in their analysis on *predictive* inferences from cause to effect. Causal relations may also be accessed in the opposite *diagnostic* direction from effect to cause. Research on causal-model theory has shown that people are capable of diagnostic inferences in trial-by-trial learning situations (see Reips & Waldmann, 2008; Waldmann *et al.*, 2006). The same causal-power assumptions underlying predictive inferences apply to diagnostic inferences. These assumptions are defaults (see Cheng, 2000, for an analysis of various relaxations of these assumptions); the first two are empirical, and may be revised in light of evidence:

(1) C and alternative causes of E influence E independently,

(2) causes in the composite background A could produce E but not prevent it,

(3) causal powers are independent of the occurrence of the causes, and

(4) E does not occur unless it is caused.

Below we illustrate how these assumptions can be applied to explain a variety of related diagnostic inferences. Consider, for example, diagnostic inferences regarding a causal structure with two causes of a common effect E: C→E ← D. How would explanation by causal powers account for the simplest diagnostic inference—the intuition that having knowledge that E has occurred, compared to the absence of such knowledge, would lead to the inference that each of the causes is more likely to have?

Similarly, how would this approach explain the intuition that given knowledge that E has occurred, the target cause C is less likely to have occurred if one now knows that an alternative cause D has occurred, compared to when one does not have such knowledge (the 'explaining away' or discounting phenomena)?

A basic case of single-effect diagnostic inference and some special variations

Here we show a causal-power explanation of an intuitive diagnostic inference from the occurrence of E to the occurrence of target cause C, namely, the intuition that the probability of C occurring given that E has occurred, $P(c|e)$, is higher than the unconditional probability of C occurring, $P(c)$. In our derivations below, c represents the event that C has occurred, and likewise for d and e with respect to cause D and effect E.

Let q_C be the generative power of C to produce E,

q_D be the generative power of D to produce E,

$e_{C\text{-}only}$ be the event that E is produced by C alone (i.e., not also by D), and

e_D be the event that E is produced by D, whether or not it is also produced by C.

By definition of conditional probability,

$$P(c \mid e) = \frac{P(c,e)}{P(e)} \tag{5}$$

Event e can be decomposed into e_D and $e_{C\text{-}only}$, two mutually exclusive events. Making use of this decomposition of e, and of causal-power assumptions, to put the right-hand-side of Equation 1 into causal power terms, one obtains:

$$\frac{P(c,e)}{P(e)} = \frac{P(d)\cdot q_D \cdot P(c) + \left[1 - P(d)\cdot q_D\right]\cdot P(c)\cdot q_C}{P(d)\cdot q_D + \left[1 - P(d)\cdot q_D\right]\cdot P(c)\cdot q_C} \tag{6}$$

The numerator shows the probabilities of the two conjunctive events – (1) C occurring and E caused by D, and (2) C occurring and E caused by C alone. The components of each conjunctive event occur independently of each other. From (5) and (6), it follows that:

$$P(c \mid e) = P(c)\cdot \frac{P(d)\cdot q_D + \left[1 - P(d)\cdot q_D\right]\cdot q_C}{P(d)\cdot q_D + P(c)\cdot \left[1 - P(d)\cdot q_D\right]\cdot q_C} \tag{7}$$

From (7) it is easy to see the relation between $P(c)$ and $P(c|e)$. In the trivial case in which $P(c) = 1$, $P(c|e)$ of course also equals 1.

When C does not always occur

But in the more interesting case, if $P(c) < 1$, Equation 7 implies that

$$P(c|e) > P(c), \tag{8}$$

thus explaining the basic diagnostic intuition that knowing that a particular effect has occurred increases the probability that its causes have occurred. As can be seen

from (7), this inequality holds regardless of the magnitude of q_D, as long as $p(d) \neq 1$; for example, in the trivial case in which C is the only cause of E (i.e., when $q_{-D} = 0$), (7) implies that given that E has occurred, the probability of C occurring is increased to 1. (See section on 'Intervening versus Observing' regarding the special case in which both $p(d)$ and q_D equal 1.)

Combining link-by-link inferences

The capacity to access causal relations in the predictive and diagnostic direction allows us to make inferences consecutively across causal networks by going from one link to the other. For example, the basic diagnostic inference just shown can be applied to the common-cause model (Fig. 20.1A) to infer the state of effect_2 from the state of effect_1 (treating effect_1 as the common effect E in the basic learning unit): First, diagnostically infer that the cause must have occurred when effect_1 is observed (i.e., $P(c|effect\_1)=1$; this is the single-cause special case in (7) in which $P(d)=0$ or $q_{c\rightarrow effect\_1} = 0$); that is, there is no causal influence from D to effect_1. Second, infer the state of effect_2 from the presence of the cause just inferred (i.e., $P(c|effect\_1) \cdot q_{c\rightarrow effect\_2}$, where $q_{c\rightarrow effect\_2}$ is the causal power of c to produce effect_2). Note that Equation 7 would similarly apply for the more general other case in which $P(d)>0$ and $q_{c\rightarrow effect\_1} > 0$; that is, there is a causal influence from D to effect_1. In the case in which D is a background cause that is constantly present, $P(d)=1$. Similar consecutive inferential steps can be derived for the chain model; for example, with respect to the causal chain C→effect_1→effect_2 (Fig. 20.1B), $P(effect\_2|c) = q_{c\rightarrow effect\_1} \cdot q_{effect\_1\rightarrow effect\_2}$. More generally, for diagnostic reasoning Equation 7 applies, and for predictive reasoning the causal powers of the relations in question apply.

It is important to note that combining these steps builds on individual links rather than any quantitatively coherent causal model representation. Note that in our derivations we did not use the Markov constraint in any of our inferential steps. Thus, it is not necessary to assume that learners use the Markov constraint in their inferences. Inferential behavior consistent with the Markov condition is in these cases a side effect of chaining the inferences, it is not an explicit part of the postulated graphical representation. For example, for the common-cause structure (Fig. 20.1A), a reasoner may or may not take the additional step of inferring the independence between effect_1 and effect_2 conditional on the common cause. If that step is not taken, say, during the initial learning of the structure when attention is focused on the learning of individual links, then a violation of that independence relation (as implied by the Markov condition when applied to the structure) will go unnoticed. Thus, more generally, although under conditions in which typical attention and memory constraints are bypassed, model construction via link-by-link causal inference will be consistent with the Markov condition, in typical situations conforming to the Markov condition for inferences regarding relationships between indirectly linked variables will depend greatly on attentional and memory factors.

Common-effect models (Fig. 20.1C) provide another interesting test case because normatively it is not permissible to chain the predictive link between cause_1 and effect, and then proceed in the diagnostic direction from the effect to cause_2 while disregarding the first link. Chaining these two inferences would erroneously predict a correlation between the two variables that are not directly linked, violating the Markov assumption as applied to Fig. 20.1C. Doing so would mean that all three models in Fig. 20.1 make the same predictions regarding the indirectly linked variables (the two effects of the common-cause model, the two causes of the common-effect model, and the cause and effect_2 in the causal chain model). However, whereas effects make each cause individually more likely, diagnosing a target cause when an alternative cause is present should make the target cause less likely than when the state of the alternative cause is unknown (i.e., explaining away). Correct inferences within a common-effect model need to consider all three event types defining causal relations: (1) target cause, (2) target effect, and (3) alternative observable and unobservable causes.

Explaining away

To explain 'explaining away' in causal-power terms, let us return to Equation 7 (assuming $P(c) < 1$). One can see that as $P(d)$ increases, $P(c|e)$ decreases, as does the difference between $P(c|e)$ and $P(c)$ (although, as just explained, $P(c|e)$ is still greater than $P(c)$ because $P(c) < 1$). In the special extreme case in which $P(d) = 1$, the left-hand-side of (7) becomes $P(c|e,d)$, and $P(c|e)$ is at its minimum (assuming unchanged causal powers). That is, the probability of C given that both E and D have occurred is less than the probability of C given E when the state of D is unknown. More generally, (7) implies that knowing that one of the two causes of an effect has occurred, relative to not knowing that, reduces the probability that the other cause has occurred (the case of discounting or explaining away):

$$P(c|e,d) < P(c|e) \tag{9}$$

We assume that people are capable of making the above inferences when their attention is brought to the relevant variables. An interesting test case for our model, however, concerns cases in which participants consecutively access the two links of a common-effect model, in the order cause_1, effect, and cause_2. People's focus on single links may mislead them into making a simple diagnostic inference from the effect to cause_2 while disregarding the previously accessed link from cause_ 1 to its effect (see below for empirical evidence). They make the mistake of inferring P(effect| cause_1) followed by P(cause_2| effect), instead of making a one-step inference regarding P(cause_2 | effect, cause_1) as just shown (see (9)).

Intervening versus observing

A final important question refers to the distinction between intervening and observing within the single-effect learning model. Observing an effect provides diagnostic evidence for its cause, whereas intervening on the effect does not. We have already elaborated on how observational diagnostic inferences should be handled. A simple model of interventional inferences would just add a causal link from the intervening agent to the manipulated variable, thus creating a new common effect structure in which the agent is a new cause (see Fig. 20.3) (see also Dawid, 2002). Let alternative cause D in this case be the added intervening agent (see Fig. 20.3). The inferences following hypothetical interventions then follow from (7), which implies explaining away when multiple causes compete for predicting a specific effect (see also Morris & Larrick, 1995). Consider the simple case in which the target cause C and the intervening agent D are the only causes of E. The analysis generalizes to more complex cases involving multiple causes of E in addition to the intervening agent.

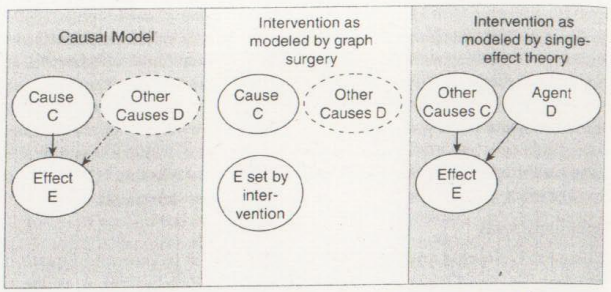

Fig. 20.3. Modeling intervention in an effect according to causal Bayes net theory and the single-effect learning theory. See text for details.

The above diagnostic inference regarding an intervention is the special case of (7) in which the intervention is always successful in producing the target effect, namely, $P(d)\cdot q_D = 1$. In that case, according to (7),

$$P(c|e) = P(c) \qquad (10).$$

That is, knowing that E has occurred does not affect the probability of C. The difference shown in our analysis between the result indicated in (10) and that in (8) explains the distinction between inferences based on intervening to obtain an effect versus merely observing the effect.

Note that when the intervention is viewed as deterministic and independent of the other cause, this analysis yields the same results as graph surgery, without removing the causal link between the usual cause and the target effect. In this special case, the manipulated effect and its usual cause become independent, which means that this cause occurs at its base rate probability.

Probabilistic Interventions

Beyond this special case the present analysis also allows for predicting the outcomes of hypothetical interventions that only probabilistically alter their target variable, or that are confounded with other events in the causal model. For example, as explained earlier, (7) shows that when the intervention is only probabilistically successful (i.e., $q_D < 1$), one can in fact infer from knowing that E has occurred (whether or not E was caused by D, the intervening variable) that the probability of C is increased relative to not knowing that E has occurred (i.e., $P(c|e) > P(c)$, Equation 8). That is, there should be no graph surgery. Thus, the classical diagnostic analysis has the advantage of greater generality.

Empirical case study 1: causal learning in rats

Although in the past decades there has been a debate about whether human causal learning can be reduced to associative learning processes or not (see Cheng, 1997; Shanks & Dickinson, 1987; Waldmann *et al.*, 2006; special issue of Learning & Behavior, 2005, Volume 33(2)), until recently most researchers have agreed that nonhuman animals are incapable of causal reasoning. Although many of these psychologists believed that even infants have the capacity for causal representations, they drew a line between human and nonhuman animals, turning causal reasoning into a uniquely human capacity similar to language. For example, Povinelli argued about chimpanzees that 'their folk physics does not suffer (as Hume would have it) from an ascription of causal concepts to events which consistently co-vary with each other' (Povinelli, 2000, p. 299; see also Tomasello & Call, 1997, for a similar argument). Gopnik and Schulz (2004, p. 375) claimed: 'The animals seem able to associate the bell ringing with food, and if they are given an opportunity to act on the bell and that action leads to food, they can replicate that action. Moreover, there may be some transfer from operant to classical conditioning. However, the animals do not seem to go directly from learning novel conditional independencies to designing a correct novel intervention.'

Blaisdell *et al.* (2006) tested whether rats are capable of causal learning and reasoning. Their goal was to show that rats distinguish between causal and spurious relations, and are capable of deriving predictions for novel actions after purely observational learning. Their experiments were modeled after a previous study on humans (Waldmann & Hagmayer, 2005). In this study participants were provided with instructions suggesting a common-cause or a causal chain model and were given data to learn about the base rates of events and about the causal strength of the causal links. In the test phase, participants were given questions regarding hypothetical observations and hypothetical interventions. For example, in one experiment they learned about a common-cause model (see Fig. 20.1A), and then were asked in what state effect_2 would be given that effect_1 was observed (observation question). The corresponding intervention question asked participants about effect_2 when effect_1 was manipulated by an external intervention. The responses showed that participants were sensitive to the distinction between observing and intervening consistent with the assumption that they had formed a causal representation of a common-cause or causal chain model. They also proved sensitive to the size of the causal strength parameters and the base rates.

In Blaisdell *et al.*'s study (2006) rats also went first through a purely observational learning phase. In their Experiment 1, rats observed three types of trials, a light followed by a tone, the light followed by food, or a click occurring simultaneously with food. These three trial types were separately presented several times during a week (see Fig. 20.4). The idea was to present rats with the individual links of a common-cause model with temporal cues suggesting the roles of potential causes and effects. We chose to separately present the link information to avoid that rats would form a model in which the two effects are directly causally instead of spuriously linked. This learning procedure was motivated by research on second-order conditioning. Yin, Barnet, and Miller (1994) have shown that with few trials rats that separately learn about two links of a chain tend to associate the first event with the last event although these two events never co-occurred. In second-order conditioning they are in fact negatively correlated. Only with many trials do rats notice the negative (inhibitory) relation. Following these findings we chose trial numbers that favored the integration of the separate links into a model in which all events are positively associated.

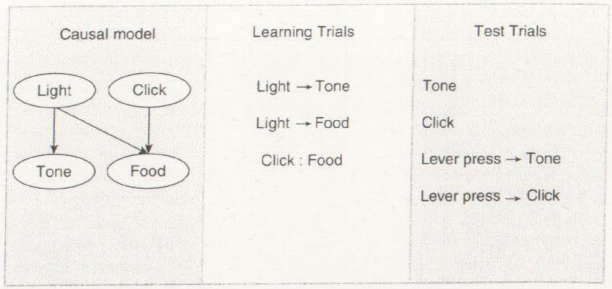

Fig. 20.4. Causal model presened to rats in Blaisdell *et al.* (2006, Experiment 1)(left). Each causal link was presented separately (→ signifies temporal order,: signifies simultaneous presentation)(middle). Test trials presented either the alternative effect of the cause of food (tone), the second cause of food (click), or these two events as a causal outcome of lever presses (click and tone were counterbalanced)(right). Rats' expectations of the presence of food were assessed by measuring their search behavior (nose poking). See text for further details.

This prediction was supported in the subsequent test phase in which rats were presented with the tone as a cue (observation test). The results showed that the tone apparently led them to believe that food was present, which was measured by the time they searched for food in a niche (i.e., nose poking). This behavior is consistent with the view that the rats accessed a common-cause model to infer from one effect (tone) to the other (food)(see Fig. 20.5). The crucial test involved a novel intervention. In this part of the test phase, a lever the rats had never seen before was introduced into the cage. (Actually there were also levers during the observation tests but pressing the levers there did not cause an event.) Whenever the rats curiously pressed the lever, the tone was presented (intervention test). Now, although tone and food had been associated by the rats in the learning phase as indicated in the observational test phase, they were less inclined to search for food after the lever presses (see Fig. 20.5). Blaisdell *et al.* (2006) viewed this behavior as evidence for the rats having formed a common-cause model in which, consistent with the Markov condition, a spurious positive correlation is implied by the two generative causal links emanating from the common cause. Whereas in the observation test rats apparently reasoned from one effect through the common cause to the second effect, they seemed to be aware of the fact that during the intervention test they and not necessarily the light were the cause of the tone. This is consistent with the view that the rats assumed that light and tone are independent during the intervention.

One could argue that the rats may have been distracted by the lever presses and therefore may have been reluctant to search for food. This possibility is ruled out by further test conditions in which the rats either observed the click or pressed the lever, which generated the click signal (see Fig. 20.4). In this condition rats expected food regardless of whether they heard the click or pressed the lever generating the click

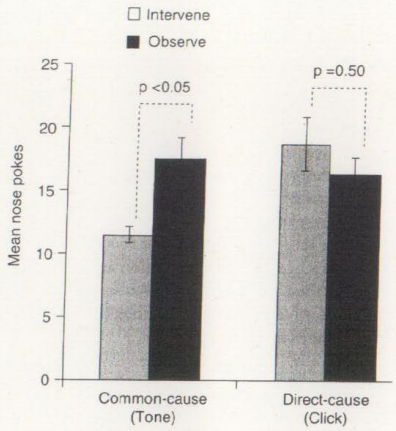

Fig. 20.5. Results of Blaisdell *et al.* (2006, Experiment 1, reproduced with permission). Rats either observed tone and intervened in click, or observed click and intervened in tone (tone and click were counterbalanced).

(see Fig. 20.5). This pattern is again consistent with the assumption that rats had formed causal knowledge. Regardless of whether a direct cause is observed or generated by an intervention the effect should occur. As an additional test, Blaisdell *et al.* (2006) presented in a second experiment a causal chain in which, again using a second-order conditioning procedure (sensory preconditioning), the tone preceded light which in turn preceded food. Consistent with a causal analysis the rats expected food regardless of whether they observed the tone or generated it with the lever. This shows again with a second-order conditioning task that the rats were not generally reluctant to expect food after a novel intervention. We will now revisit the results of Blaisdell *et al.* (2006) discussing them in greater detail in the context of the three models.

Associative theory

Although previous research on second-order conditioning has focused on chain-like structures and not common-cause models, the findings are consistent with second-order conditioning in both experiments. However, why second-order conditioning occurs is not entirely clear, especially because it seems to be dependent on trial number (Yin *et al.*, 1994). According to associative theories, rats should associate light with tone, and light with food. If this is represented in the typical one-layer network with one cue (light) and two independent outcomes (tone, food), two positive associative weights should be formed for either link. Without any further learning these outcomes would indeed be correlated, which is consistent with the findings in the observational test phase. However, this prediction only holds if it is assumed that the inhibitory relation between the two outcomes tone and food is not encoded, at least with few trials. Thus, according to the associative view, rats might associate tone with food through second-order conditioning when the additional assumption is made

that no associations between outcomes will be learned. However, the associative view breaks down when the intervention test is considered. It cannot explain why, given that rats associate tone with food, they nonetheless did not expect food when their actions caused the tones.

It is important to note that acquisition-based theories (e.g., Rescorla & Wagner, 1972) do not predict that lever presses compete with light as an explanation of tones because the light-tone trials and the lever press-tone trials were separately presented, thus preventing cue competition.[1] Another possible argument might be that the intervention test is in fact a novel instrumental conditioning task. Thus, rats may not expect food in Experiment 1 because the instrumental action is novel so that they had not formed any associations between lever presses, tone and food. However, this explanation is contradicted by the direct cause (click) condition and by the causal chain experiment (Experiment 2). In these conditions, interventions and observations led to equal amounts of search for food. In sum, the results by Blaisdell *et al.* (2006) are inconsistent with current associative theory.

Causal Bayes nets

Causal Bayes net theory (Gopnik *et al.*, 2004) can be applied to both the learning phases and the testing phases of Blaisdell *et al.*'s (2006) experiments. It is obvious that a bottom-up constraint-based learning algorithm is incapable of explaining the results of the learning phase, even when temporal order cues are used that aid the induction process. The temporal order of events suggests that light is a potential cause of tone and of food. However, the learning patterns are inconsistent with a common-cause or causal chain model in which the Markov condition holds. For example, in Experiment 1 rats observe the patterns light-tone-absence of food (P(t.l.~f)) or light-food-absence of tone (P(l.f.~t)), along with additional click-food trials. According to the Markov condition, P(t|l.f) should be equal to P(t|l.~f). But the first probability is zero, the second probability is one, which clearly violates the Markov condition. The same problem arises for the causal chain condition (Experiment 2) in which tone is negatively correlated with food.

An alternative to constraint-based learning might be a Bayesian algorithm which assigns prior probabilities to all possible models. It might be possible to develop a model, which makes use of temporal order cues (i.e., light is the potential cause) and assigns very high prior probabilities to common-cause and causal chain models that honor the Markov condition. This model might also predict Yin *et al.*'s (1994) finding that the negative correlation in the chain structure only becomes salient after many trials. However, this model is computationally very demanding. Bayesian algorithms typically require many learning trials to converge, more than are usually presented in experiments with humans and nonhuman animals (see Tenenbaum & Griffiths, 2003). Moreover, it is post hoc: The strong assumption needs to be made that rats represent simple causal models as priors that are strong enough to override the violation of the Markov condition for those models (e.g., the common-cause model in Figure 1A rather than one with added inhibitory links between the two effects). Thus far, there is no independent evidence for this claim.

[1] Matute and Pineño (1998) and Escobar, Matute, and Miller (2001) found evidence that cues that had been paired separately to a common outcome can compete, but in recent studies we (Leising, Wong, Stahlman, Waldmann, & Blaisdell, in press) have demonstrated that even this associative mechanism cannot account for the effects reported by Blaisdell *et al.* (2006).

The results of the test phase are indeed consistent with the assumption that rats formed and accessed a Bayesian common-cause or causal chain network in Experiments 1 and 2 (Blaisdell *et al.*, 2006). However, the causal Bayes net account suffers from the problem that it is unclear how these models were induced from the learning data. Thus, there is a gap between learning and testing that currently cannot easily be filled by causal Bayes net theory.

Single-effect learning model

According to this theory simultaneously considering complex patterns of events involving multiple effects is too demanding for rats (and possibly also often for humans). Instead we assume that rats focus their attention on single effects, as mentioned earlier. According to this model, in the learning phase rats should either focus on the light-tone, or the light-food relation. (There is also the click-food link which we will ignore in this section.) According to the model temporal cues are used to distinguish potential causes (e.g., light) from potential effects (tone, food). When the light cue is present, rats may learn to expect both tone and food. However, the focus on single effects will lead the rats to update with respect to one effect at a time. Hence, once tone or food is present, they should focus on learning the link that leads to the present effect, and ignore the second link.

This model therefore explains how rats learn about two separate links that happen to share a common element without assuming that they use information about how the indirectly linked elements are related to each other. More specifically, it need not be assumed that the rats make the Markov assumption and represent the three events as part of a Bayesian common-cause model.

How does the single-effect learning model explain the behavior in the test phase? During the observation tests rats hear tones as cues. The tones lead them to diagnostically infer the light. Then they proceed in the predictive direction from light to food. The link-by-link inferences according to this model explain why the rats expect food although in the learning phase tone and food are negatively correlated. Since the links overlap in the light event, rats are capable of making an inference across the network. It is important to note that there is no need to assume that rats represent a coherent common-cause model obeying the Markov assumption. Inferences consistent with the Markov condition are a side effect of chaining separately represented links in the common-cause and the chain models; the Markov condition is neither part of the representation of available information nor of the computational steps involved in the inference processes.

One curious finding is that in the test phase rats infer food from a tone cue although light—the common cause—is absent. This creates an inconsistency between (1) the inference steps going from tone to light and then to food, and (2) the observed information. One possible explanation might be that in the test phase the rats focus on the target cue but do not check whether the state of other events outside their

attentional focus is consistent with their predictive steps. This may be particularly plausible to rats in Experiment 1 because the effect of lever presses—the tones—should occur *after* their usual cause, the light, so that the absence of the light might not be salient.[2] Both the learning and the test phases show that rats do not seem to treat absent events outside their attentional focus as informative (see below for further elaborations).

The second test condition involves interventions. According to the single-effect learning model interventions are represented as external causes. In Blaisdell *et al.*'s (2006) experiments this means that the lever presses should be represented as an additional cause of tones, which turns the tones into a common effect of light and lever presses. In the test phase lever presses deterministically cause tones. Although lever presses cause tones only in the absence of light, a plausible assumption is that human and nonhuman animals tend to view their arbitrary interventions as independent of alternative causes. This should lead to a discounting of the cause (light) and hence to a lowering of the expectation of food.

Under the condition that lever presses are viewed as deterministic and independent causes, it is possible to infer the probability of light. We have already shown that in this special case there should be complete explaining away of the tone by the lever press, with light to be expected to occur at its base rate (see (10)). Again the inference is modeled as a chaining of individual links. First, as explained by comparing Equation 10 with Equation 8, the presence of the tone after an intervention, relative to merely observing the tone, should lead to a lowered expectation of light; second, the expectation of light in turn should lead to a lowered expectation of food. This prediction is equivalent to the assumption of graph surgery in a causal Bayes model but it neither requires the assumption that the common cause renders its effects conditionally independent (i.e., Markov condition) nor the deletions of preexisting causal relations. Traditional explaining away will generate the same prediction as graph surgery, and additionally has the advantage of being the more general account (e.g., probabilistic interventions; conditional or confounded interventions) (see also Dawid, 2002).

The single-effect learning model in its present version predicts inferences conforming to a common-cause model with positively correlated effects regardless of whether the effects were positively or negatively correlated in the learning phase, if the relation between the two effects is unnoticed. The positive correlation is generated as a consequence of sequential access to individual causal links, rather than being directly acquired during the learning phase. If increasing the number of trials increases the chance that the relation between the two effects is noticed, however, this model would also be consistent with Yin *et al.*'s (1994) finding that with many trials rats become aware of the negative correlation of the indirectly related events. The assumption that the Markov condition is merely a consequence of link-by-link inference under propitious

[2] This assumption is actually less plausible in the chain condition of Experiment 2 in which light follows tone. Interestingly, Blaisdell *et al.* (2006) needed to hide the light behind a salient cover to obtain second-order conditioning. Hiding the light makes its absence ambiguous, it could be absent but it could also be merely invisible.

circumstances rather than a constraint in inference explains why support for the role of that condition can vary from situation to situation. Once the salient individual causal relations are learned, rats may become capable of attending to less salient relations. According to this view, learning is dependent. At the beginning of the learning phase all attention needs to be devoted to picking up single cause-effect contingencies. Once learning is stabilized, this may free attention limitations.

In summary, the single-effect learning model provides the most parsimonious account of the three theories of Blaisdell *et al.*'s (2006) results. The model implies that rats indeed learn and reason about causal relations in a sense that is inconsistent with current associative theories. Moreover, it demonstrates that computations using the Markov condition underlying causal Bayes nets are not necessary to account for the data. Furthermore, the less computationally demanding causal inferences in the single-effect model can explain what Bayes net theory fails to explain. Thus, the model is an example of a minimal rational model.

Empirical case study 2: combining causal relations

We rarely acquire knowledge about complex causal models all at once. Often we only learn about fragments of causal knowledge, which we later combine to more complex causal networks (see also Lagnado *et al.*, 2007; Waldmann, in press). For example, we may learn that we tend to get a stomach ache when we eat sushi or when we take an aspirin. We may never have taken aspirin with sushi but still will have a hunch what the effect on our stomach might be.

Hagmayer and Waldmann (2000; in preparation) have investigated the question of how people combine individually learned causal relations (see also Ahn & Dennis, 2000; Perales *et al.*, 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 instructions about the underlying causal model differed in the contrasted 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 (Fig. 20.6 only shows the results for the conditions in which strength was strong). Note that participants, like the subjects in Blaisdell *et al.*'s (2006) rat studies, learned about each causal link individually. However, participants were told that the events involved in the two causal relations were studied at two universities, which invites the inference that the second cause or effect currently not presented is not necessarily absent but simply not measured.

The main goal of the study was to test what predictions people would make about the correlation between the events that were presented in separate trials. A correlation should be expected between the two substances when they were effects of a Bayesian common-cause model that honors the Markov condition with the size of the correlation being dependent on the size of causal strength of the causal links. By contrast,

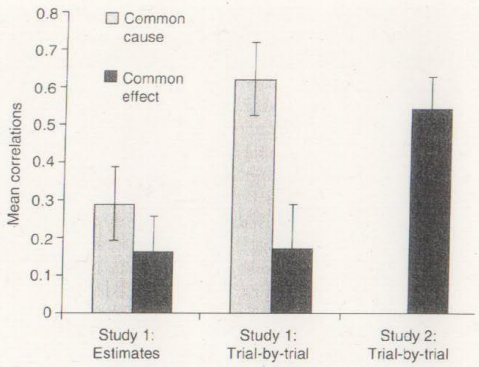

Fig. 20.6. Selected results from Hagmayer and Waldmann (2000; in prep.). In study 1, participants first made multiple trial-by-trial predictions of the two effects (common-cause model) or the two causes (common-effect model) based on the assumed presence or absence of the cause or the effect, respectively (middle). Subsequently they provided conditional frequency judgments concerning the two effects (common-cause model) or the two causes (common-effect model)(i.e., estimates, left). In the second study (right), participants were requested to predict, again across several trials, first the common effect on the basis of information about cause_1, and then to diagnose cause_2. The graph shows the mean correlations (and standard errors) that were derived from participants' conditional frequency estimates or the patterns generated in the trial-by-trial judgments.

two causes of a common effect should be independent regardless of the strength of the causal relations.

To test this knowledge, participants were given two different tasks in Study 1: 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 each of the two substances was present or absent. Thus, in the common-cause condition 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 each 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 the expected implied correlations. The second task asked participants directly to estimate the conditional frequency of the second substance in a set of trials given that the first substance was either always present or always absent: These two estimates were combined to calculate the inferred correlations between the substances.

The results of this and other experiments show little sensitivity to the differences between common-cause and common-effect models in the conditional frequency estimations. Although some basic explicit knowledge cannot be ruled out (see Perales *et al.*, 2004), Hagmayer and Waldmann's (2000, in preparation) experiments show that people exhibit little awareness of the relation between the causal strength of the links and the implied spurious correlation. By contrast, the task in which participants made trial-by-trial predictions corresponded remarkably well to the predictions entailed by the contrasted causal models. Whereas a spurious correlation was predicted in the common-cause condition, the predicted correlation stayed close to zero in the common-effect condition (see Fig. 20.6).

In further experiments Hagmayer and Waldmann (in preparation) followed up on the online trial-by-trial prediction measure, which in the first experiment yielded results corresponding to Bayesian common-cause and common-effect models. As it turns out, this effect can only be found in a task in which the two substances were predicted simultaneously. In Study 2 participants again learned the individual links of a common-effect model. Now one of the tasks was to first predict the effect based on one cause, and then make inferences about the other cause. In this study people's inferences exhibited a spurious correlation between the causes of a common effect, similar to what they predicted for the effects of a common cause. We will again use these studies to evaluate the different theoretical accounts.

Associative theory

Associative theories could be used to explain why people generate a correlation between multiple outcomes of a common cue, as in the common-cause condition of the first study of Hagmayer and Waldmann (2000; in preparation). Within a one-layer network multiple outcomes of a common cue should be correlated. This theory may also explain the correlation of multiple causes of a common effect found in the second study as an instance of second-order learning. However, this model is refuted by the finding in Study 1 that, when the cue represented a common effect and the outcomes alternative causes, with identical learning input participants did not generate a correlation between the causes in their trial-by-trial predictions. This result clearly supports causal over associative learning theories. It also demonstrates people's capacity to separate the representation of causes and effects from the representation of temporally ordered cues and outcomes: Although effect information temporally preceded cause information in the common-effect condition of Study 1, participants correctly induced a common-effect model. This finding adds to the substantial number of studies that have shown that humans are capable of disentangling temporal order from causal order (see Lagnado *et al.*, 2007; Waldmann, 1996; Waldmann *et al.*, 2006, for overviews).

Causal Bayes nets

Unlike in Blaisdell *et al.*'s (2006) studies, the learning phases in these experiments do not violate the Markov condition. Learners may learn about each link separately, and make assumptions about the probable state of the third event that currently was not measured. With the aid of the instructions that suggested which events were causes and effects, it is possible to learn a common-cause and common-effect model that is

consistent with the Markov condition by updating causal strength estimates within the instructed model. Causal Bayes net theory can also explain why people in Study 1 generate correlations in the common-cause condition, but not in the common-effect condition. These inferences also fall out of the assumption that people represent and access Bayesian causal models that honor the Markov condition. However, causal Bayes net theory fails to explain why people only conformed to the predictions of this theory when they made trial-by-trial online predictions but not when they made conditional frequency judgments. Causal Bayes net theory, being derived from a normative theory of theory discovery, predicts behavior conforming to their normative prescriptions regardless of the way the data are presented or the test questions are asked. Consequently, causal Bayes net theory also fails to explain why people generate a correlation between alternative causes of a common effect when asked about the two links consecutively, as in Study 2. This inference clearly violates the assumptions underlying common-effect models.

Single-effect learning theory

For the learning phase we assume that participants, like the rats in Blaisdell *et al.* (2006), update each link individually. Unlike associative theories, this theory distinguishes between causes and effects, and can therefore capture the difference between diagnostic effect-cause and predictive cause-effect learning. Information is stored in the weight and direction of individual links in a causal network. Due to its focus on individual causal effects, the theory explains why participants in the first study did not have explicit knowledge (i.e., conditional frequency estimates) about the structural implications of common-cause versus common-effect models. Moreover, this theory predicts that learners correctly generated the correlations implied by the different causal models. In the common-effect condition in Study 1, learners were presented with a common-effect model with a single effect and alternative causes. The task was to simultaneously diagnose the alternative causes on the basis of specific effect tokens, which served as cues. According to our analysis, learners should focus on individual effects and be aware of the 'discounting' of a cause by alternative observable and unobservable causes (9). Thus, learners should be reluctant to infer the presence of multiple causes, when one cause sufficiently explains the effect. In contrast, in the common-cause condition learners should generate predictions by focusing on one effect after the other. This strategy would generate correlated effects although participants may not become aware of the correlation. Interestingly, this prediction is supported by the fact that learners did not show any awareness of the effect correlation in the conditional frequency judgments although these judgments followed the online trial-by-trial generation phase in Study 1 (see Fig. 20.6).

The results of Study 2 can also be explained by the single-effect learning model. In this study participants were led to consecutively access the two causal links. Diagnosing a cause from a single effect may be ambiguous for learners. If they simply focus on the state of the effect as unconditional information and compute the likelihood of the second presented cause, the effect should provide positive diagnostic support for the cause. Only if learners consider that the effect token can be produced by cause_1 should the diagnostic inference to effect_2 be lowered. However, this inference requires considering all three events at once. In Study 1, the task highlighted the potential competition of the two causes by having learners diagnose them both at once, whereas in Study 2 the consecutive nature of the task may have led participants to access the second link while disregarding the first link, as they would do in a causal chain or common-cause situation. Thus, in sum, the single-effect learning account provides the broadest and at the same time simplest model for the data.

Note that in Hagmayer and Waldmann's (2000, in preparation) studies cover stories were used that encouraged the assumption that the currently non-observed second effect may actually be present but is just not shown. Thus, participants knew that the fact that they only see one effect at a specific learning trial does not necessarily mean that the second effect is absent. We did not present participants with a learning phase in which a common cause generates negatively correlated effects, as in Blaisdell *et al.*'s (2006) study. Consistent with our speculation on the role of attention in rats' learning, we predict that human learners, due to their greater attention span, should in fact become aware of the negative correlation between the effects of a common cause much earlier if this information is saliently presented.

Nevertheless, we expect that learning about a common cause with negatively correlated effects is actually a more difficult task than learning about standard common-cause models. Not only the single-effect learning model, but also causal Bayes net models cannot easily represent such causal structures. A typical solution, to add an inhibitory link between the effects, seems rarely plausible as a description of the underlying causal mechanisms. One example for such an atypical common-cause model would be a situation in which a beam emitting x-rays could be spatially focused on two different locations. Another example, which is suggested by Cartwright (1989), is the negative dependency that arises when one has a limited amount of money for buying meat and vegetables in a grocery store so that the first (presumably independent) decision limits the second. In both scenarios, a simple representation of the common cause as independently influencing two effects is inappropriate. A deeper re-representation is required that reflects the underlying mechanism or capacity limitations. Instead of inferring the state of the common cause from one effect, and using this state to make further inferences, different hidden properties of the cause need to be induced and used for inferences in these more complex scenarios (see Rehder & Burnett, 2005, for relevant findings). In sum, we assume that more complex structures are indeed learnable but that they require some extra effort. The initial bias of learners may still be the simpler inferences afforded by the proposed single-effect learning model.

Conclusion

Our test cases demonstrated the value of rational models while at the same time adhering to the traditional standards of empirical theory testing. Like other psychological theories, rational models can be tested against each other. We also demonstrated the usefulness of the heuristic to search for minimal rational models. Our two test cases showed that both human and nonhuman animals go beyond the information given by inferring unobservable causal processes on the basis of observable data.

Thus, Hume's view that we are restricted to observable covariations is refuted. Both causal Bayes net theory and our single-effect learning theory provide a rational account of causal learning. Both theories claim that the goal of causal learning is to adapt to the causal texture of the world, but based on different assumptions about cognitive capacities their answers are different.

The search for minimality highlights the deficits of causal Bayes net theory. The problem with this theory is that it is overly powerful. It was originally developed as a normative tool and is therefore developed to yield normative answers in all possible circumstances. Thus, it is ill-prepared to account for failures and strategy-based restrictions of human and nonhuman learning, and it overestimates the complexity of reasoning in biological systems. Causal Bayes net theory makes strong assumptions about the structure of causal network (i.e., Markov condition) that may be methodologically useful but may not represent what people and animals actually believe (see also Cartwright, 2001, 2004). Thus far, there is little evidence that people assume the Markov condition when reasoning about causal models (see Rehder & Burnett, 2005). Although it is possible to construct a more complex Bayes net with hidden nodes that predicts violations of the Markov condition in inference patterns (e.g., Rehder & Burrett, 2005) it remains to be seen whether the structural constraints underlying these models prove plausible as accounts of reasoning and learning. Our single-effect learning theory showed that it is possible to model reasoning regarding causal models without using the Markov condition as a constraint. Depending on the task requirements, such as the way the knowledge is accessed however, the predictions may or may not conform to the Markov condition.

Although the empirical evidence we presented favored the single-effect learning model, it is too early to decide whether it will prove superior in other learning scenarios as well. We have already discussed the problems that arise with unusual causal structures, such as common-cause models with negatively correlated effects. Moreover, we have only discussed studies in which learners were presented with individual causal relations in the learning phase. This may have favored strategies consistent with our model. Future research will have to test the generality of this theory, for example, in learning situations in which multiple effects are presented simultaneously.

We presented two sets of studies to illustrate how rational models can be tested. Although in early stages of research there may be only a single rational model, the possible tradeoffs between different factors entering rational model construction make it likely that in more advanced stages there will be competing theories. Theories can best be evaluated when they are compared with each other. Each of the three theories we have selected is supported by empirical evidence, which the researchers endorsing the theories discuss in the articles presenting the theories. Fitting data to individual models is not a very convincing way to test theories (see Roberts & Pashler, 2000). Most of the time there is evidence supporting the theories, and data contradicting the theory can often easily be explained away as noise, performance factors, or as results which do not fall under the scope of the theories. A more promising strategy is to test specific competing theories against each other. Unlike what Anderson (1990) has proposed we believe that all the relevant psychological evidence that can be found should be brought to bear on the models. Taking into account all available psychological data reduces rather than increases the possibility of indeterminacy. Currently indeterminacy seems more like a theoretical than a practical threat anyhow. We are not aware of any cases in psychology in which theories make identical predictions in all situations. Should this case occur, then there is indeed no way to decide between the equivalent theories except on the basis of other criteria, such as simplicity. However, we are far from reaching the luxurious state of having to choose between equivalent theories.

References

Ahn, W.-K., & Dennis, M. J. (2000). Induction of causal chains. In L. R. Gleitman & A. K. Joshi (Hrsg.), *Proceedings of the Twenty-Second Annual Conference of the Cognitive Science Society*. Mawah, NJ: Erlbaum.

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

Allan, L. G. (1993). Human contingency judgments: Rule-based or associative? *Psychological Bulletin, 114*, 435–448.

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

Blaisdell, A. P., Sawa, K., Leising, K. J., & Waldmann, M. R. (2006). Causal reasoning in rats. *Science, 311*, 1020–1022.

Buehner, M. J., & Cheng, P. W. (2005). Causal learning. In K. J. Holyoak & R. G. Morrison (Eds.), *Cambridge handbook of thinking and reasoning* (pp. 143–168). Cambridge, UK: Cambridge University Press.

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.

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

Cartwright, N. (2001). What is wrong with Bayes nets? *The Monist, 84*, 242–264.

Cartwright, N. (2004). Causation: One word, many things. *Philosophy of Science, 71*, 805–819.

Chater, N., & Oaksford, M. (2004). Rationality, rational analysis and human reasoning. In K. Manktelow & M. C. Chung (Eds.), *Psychology of reasoning: Theoretical and historical perspectives* (pp. 43–74). Hove, Sussex: Psychology Press.

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

Cheng, P.W. (2000). Causality in the mind: Estimating contextual and conjunctive causal power. In F. Keil & R. Wilson (Eds.), *Cognition and explanation* (pp. 227–253). Cambridge: MIT Press.

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

Dawid, A. P. (2002). Influence diagrams for causal modelling and inference. *International Statistical Review, 70*, 161–189.

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

Escobar, M., Matute, H., & Miller, R. R. (2001). Cues trained apart compete for behavioral control in rats: Convergence with the associative interference literature. *Journal of Experimental Psychology: General, 130*, 97–115.

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

Glymour, C. (1980). *Theory and evidence*. Princeton: Princeton University Press.

Goodman, N. *Fact, fiction, and forecast* (4th ed.). Cambridge, MA. Harvard University Press.

Gopnik, A., & Schulz, L. E. (Eds.)(2007). *Causal learning: Psychology, philosophy, and computation*. Oxford University Press.

Gopnik, A., & Schulz, L. E. (2004). Mechanisms of theory-formation in young children. *Trends in Cognitive Sciences, 8*, 371–377.

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.

Griffiths, T. L., & Tenenbaum, J. B. (2005). Structure and strength in causal induction. *Cognitive Psychology, 51*, 285–386.

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* (pp. 214–219). Mahwah, NJ: Erlbaum.

Hagmayer, Y., & Waldmann, M. R. (in preparation). *Integrating fragments of causal models: Theory- and simulation-based strategies*. Unpublished manuscript.

Hagmayer, Y., Sloman, S. A., Lagnado, D. A., & Waldmann, M. R (2007). Causal reasoning through intervention. In A. Gopnik & L. E. Schultz (Eds.), *Causal learning: Psychology, philosophy, and computation* (pp. 86–100). Oxford University Press.

Hume, D. (1748/1977). *An enquiry concerning human understanding*. Indianapolis: Hackett Publishing Company.

Kant, I. (1781/1965). *Critique of pure reason*. Macmillan, London.

Lagnado, D. A., Waldmann, M. A., Hagmayer, Y., & Sloman, S. A. (2007). Beyond covariation. Cues to causal structure. In A. Gopnik & L. E. Schultz (Eds.), *Causal learning: Psychology, philosophy, and computation* (pp. 154–172). Oxford University Press.

Leising, K., Wong, J., Stahlmann, W. D., Waldmann, M. R., & Blaisdell, A. P. (in press). The special status of actions in causal reasoning in rats. *Journal of Experimental Psychology: General*.

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

Liljeholm, M., & Cheng, P. W. (2007). When is a cause the 'same'? Coherent generalization across contexts. *Psychological Science, 18*, 1014–1021.

Mackie, J. L. (1974). *The cement of the universe. A study of causation*. Oxford: Clarendon Press.

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

Matute, H. & Pineño, O. (1998). Stimulus competition in the absence of compound conditioning. *Animal Learning & Behavior, 26*, 3–14.

Morris, M. W., & Larrick, R. P. (1995). When one cause casts doubt on another: A normative analysis of discounting in causal attribution. *Psychological Review, 102*, 331–355.

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

Oaksford, M., & Chater, N. (2007). *Bayesian rationality*. Oxford: Oxford University Press.

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

Pearl, J. (2000). *Causality: Models, reasoning, and inference*. Cambridge, MA: 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.

Povinelli, D. J. (2000). *Folk physics for apes*. Oxford, England: Oxford University Press.

Quine, W. V. O. (1960). *Word and object*. Cambridge, Mass.: MIT Press, *55* 9–22.

Rehder, B., & Burnett, R. (2005). Feature inference and the causal structure of categories. *Cognitive Psychology, 50*, 264–314.

Reips, U.-D., & Waldmann, M. R. (2008). When learning order affects sensitivity to base rates: Challenges for theories of causal learning. *Experimental Psychology, 55*, 9–22.

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

Roberts, S., & Pashler, H. (2000). How persuasive is a good fit? A comment on theory testing. *Psychological Review, 107*, 358–367.

Salmon, W. C. (1980). Probabilistic causality. *Pacific Philosophical Quarterly, 61*, 50–74.

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

Shafer, G. (1996). *The art of causal conjecture*. Cambridge, MA: The MIT Press.

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

Shultz, T. R. (1982). Rules of causal attribution. *Monographs of the Society for Research in Child Development*, 47 (No.1).

Skyrms, B. (2000). *Choice & chance: An introduction to inductive logic* (4th ed.). Belmont, CA: Wadsworth/Thomson Learning.

Sloman, S. A. (2005) *Causal models: How we think about the world and its alternatives*. Oxford: Oxford University Press.

Sneed, J. D. (1971). *The logical structure of mathematical physics*. Dordrecht: Reidel.

Spellman, B. A. (1996). Acting as intuitive scientists: Contingency judgments are made while controlling for alternative potential causes. *Psychological Science, 7*, 337–342.

Spirtes, P., Glymour, C., & Scheines, P. (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.

Suppes, P. (1970). *A probabilistic theory of causality*. Amsterdam: North Holland.

Tenenbaum, J. B., & Griffiths, T. L. (2003). Theory-based causal inference. *Advances in Neural Information Processing Systems, 15*, 35–42.

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

Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and. probability. *Cognitive Psychology, 5*, 207–232.

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: Causal learning* (pp. 47–88). San Diego: Academic 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. (2001). Predictive versus diagnostic causal learning: Evidence from an overshadowing paradigm. *Psychological Bulletin & Review, 8*, 600–608.

Waldmann, M. R. (in press). Combining versus analyzing multiple causes: How domain assumptions and task context affect integration rules. *Cognitive Science*.

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. (2005). Seeing vs. doing: Two modes of accessing causal knowledge. *Journal of Experimental Psychology: Learning memory and Cognition, 31*, 216–227.

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., Hagmayer, Y, & Blaisdell, A. P. (2006). Beyond the information given: Causal models in learning and reasoning. *Current Directions in Psychological Science, 15*, 307–311.

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.

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

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

Yin, H., Barnet R. C., & Miller R. R. (1994). Second-order conditioning and Pavlovian conditioned inhibition: Operational similarities and differences. *Journal of Experimental Psychology: Animal Behavior Processes*, 20, 419–428.

Chapter 21

The value of rational analysis: an assessment of causal reasoning and learning

Steven Sloman and Philip M. Fernbach

Brown University, Providence, RI, USA

Our goal in this chapter is a rational analysis of human causal reasoning and learning. We take a rational analysis to be an assessment of the fit between data and a certain kind of model (Danks's chapter offers a more multi-faceted view of rational analysis). In the rational analysis tradition of Anderson (1990) and Oaksford and Chater (1998; in press), the term 'rational' has come to have three different meanings that vary in normative force. The first section of this chapter will be devoted to explicating these different meanings and evaluating their usefulness. The second section will apply these interpretations to assess the rationality of causal reasoning and learning.

The value of a rational model

In the rational analysis tradition, 'rational model' and 'computational model' tend to be used synonymously (e.g., Griffiths *et al.*, in press). Danks (Chapter 3, this volume) challenges this equation. According to Marr (1982), who introduced the computational level of description, a computational model describes the goal of a computation, why it is appropriate, and the logic of the strategy by which it can be carried out. What is missing from Marr's analysis is what determines the computation. Is it determined through an analysis of the task or must the analyst first observe what computation is actually being performed before engaging in a computational analysis? This is the critical question in determining whether or not a computation is 'rational.' Here are three different senses of 'rational model':

Normative model

This sense of rational model has its origins in Savage's (1972) analysis of subjective probability, a concept whose influence in psychology is primarily due to Kahneman and Tversky (1982). A rational model in this sense is a representation of the best way to perform a task. Given some goal, a normative model dictates what is necessary to achieve that goal. For instance, in the context of causal reasoning, if a machine is broken, a normative model might dictate the most cost-effective action to fix it.