To appear in: D. Reisberg (Ed.), *Oxford Handbook of Cognitive Psychology*. New York: Oxford University Press.

Causal Reasoning

Michael R. Waldmann and York Hagmayer

Department of Psychology, University of Göttingen, Germany

Abstract

Causal reasoning belongs to our most central cognitive competencies. Causal knowledge is used as the basis of predictions and diagnoses, categorization, action planning, decision making and problem solving. Whereas philosophers have analyzed causal reasoning for many centuries, psychologist have for a long time preferred to view causal reasoning and learning as special cases of domain-general competencies, such as logical reasoning or associative learning. The present chapter gives an overview of recent research about causal reasoning. It discusses competing theories, and contrasts domain-general accounts with theories that model causal reasoning and learning as attempts to make inferences about stable hidden causal processes.

Introduction[1]

Causal reasoning belongs to one of our most central cognitive competencies, which enable us to adapt to our world. Causal knowledge allows us to predict future events, or diagnose the causes for observed facts. We plan actions and solve problems using knowledge about cause-effect relations. Without our ability to discover and empirically test causal theories, we would not have made progress in various empirical sciences, such as physics, medicine, biology, or psychology, and would not have been able to invent various technologies that changed our lives. Yet causal reasoning has been curiously absent from mainstream cognitive psychology for many decades. The situation has

[1] We would like to thank M. Buehner, D. Lagnado, and K. Holyoak for helpful comments.

dramatically changed in the past two decades with more and more research devoted to causal reasoning and causal learning.

The goal of the present chapter is to present an overview of various theoretical paradigms studying causal reasoning. Whereas cognitive psychology has for a long time neglected this topic, causality and causal reasoning has remained one of the central themes of philosophy throughout its history. In fact, psychological theories of causal reasoning have been greatly inspired by philosophical accounts. So far there is no dominant overarching theory of causality or causal reasoning that would serve as an organizational foundation for presenting research on different reasoning tasks (e.g., prediction, diagnosis, inductive reasoning, decision making). We have therefore decided to structure the research according to the postulated concept of causality, which, at this stage of research, seemed more natural than organizing the chapter around different types of reasoning tasks.

Causal Reasoning without Causation: Associationist, Logical, and Probabilistic Theories

One of the reasons for the neglect of causality in cognitive psychology may have been the widespread skepticism about the reality of causation in philosophy and science. An outspoken proponent of this critical view, Bertrand Russell, stated: "The law of causation,… is a relic of a bygone age, surviving, like the monarchy, only because it is erroneously supposed to do no harm" (1912/1992, p. 193). This skepticism has in the meantime been questioned because a closer look at how scientists actually reason has revealed that causal constructs are central for scientific discovery (see Cartwright, 1999; Pearl, 2000). Nevertheless, many psychological theories have tried to explain causal reasoning with theories that do not contain references to specifically causal concepts.

The following three sections will present three classes of theories which try to reduce causal reasoning to domain-general non-causal reasoning. Although some of these

theories acknowledge the difference between causes and effects, which can be viewed as a first step in the direction of a causal theory, the postulated reasoning and learning mechanisms merely capture covariation information or logical relations without expressing causal notions.

(1) Associative Theories

Skepticism about the usefulness of the concept of causality can in psychology be traced back to the critical analyses of the philosopher David Hume (e.g., Hume, 1748/1977). Hume reflected about situations in which he observed causes and effects, and did not detect any empirical features that might correspond to evidence for hidden causal powers, which necessitate effects. What he found instead was spatiotemporally ordered successions of events. So, why do we believe in causation? His answer was that our impression of causation was merely an illusion derived from observed associations between event pairs. Contemporary learning theorists have adopted Hume's empiricist approach to causal learning. Associations derived from spatiotemporally connected events, such as through Pavlovian and instrumental conditioning, serve in these theories as the basis for causal predictions (e.g., Allan, 1993; Shanks & Dickinson, 1987; see López & Shanks, 2008, for a recent overview). Although associative theories claim to model causal reasoning, there is actually no place for causation in these theories, regardless of the variant. All associative theories divide learning events in two classes, cues and outcomes, which are distinguished on the basis of temporal order. Cues represent events that are experienced first in a learning context, and which trigger internal representations of outcomes based on the strength of the associations, which reflect the degree of covariation between the learning events.

Associative theories serve as an interesting contrast to causal reasoning. Many reasoning tasks can be successfully solved on the basis of associations, and do not require causal knowledge that goes beyond covariations. Covariation information can often be

used to make successful predictions and diagnoses, and can be employed for action planning and decision making. Moreover, covariation information used in modern associative theories is often quite sophisticated. For example, the popular Rescorla-Wagner (1972) model computes associative weights for multiple cues of a single outcome which are (under specific conditions) formally equivalent to partial regression weights in multiple regression analysis. Thus associative weights do not simply reflect simple unconditional covariations, they take into account the predictive contribution of competing cues (i.e., cue competition).

An example of cue competition, which has also been adopted in research on causal reasoning, is the blocking paradigm in which in a first learning phase a particular cause A is paired with an effect (e.g., Beckers, De Houwer, Pineño, & Miller, 2005; De Houwer & Beckers, 2003; Chapman & Robbins, 1990; Shanks, 1985; Sobel, Tenenbaum, & Gopnik, 2004; Waldmann & Holyoak, 1992). In a second phase, A is redundantly paired with a novel cause, B, and the compound of causes A and B are now followed by the effect. Although both A and B individually covary with the effect, participants view A as a cause but tend to be uncertain about whether B is a cause. This can be interpreted as evidence for cue competition. Once A is known to be a cause, B does not add anything to the predictability of the effect event. Blocking in causal reasoning is just one example of how associative theories proved useful as models of causal reasoning. Numerous other empirical findings (e.g., acquisition curves; trial order effects; sensitivity to contingency) also seem to support the view that causal reasoning does not need the concept of causation, and can be reduced to a sophisticated form of covariation learning and associative reasoning.

Despite the success of associative theories, numerous studies in the past two decades have shown that humans are sensitive to aspects of causation that cannot be reduced to covariation information. The first demonstration in which causal and

associative theories were directly pitted against each other comes from Waldmann and Holyoak's work on causal model theory who have shown that learners distinguish between cues that represent causes of outcomes (i.e., predictive learning), and cues that represent effects of outcomes (i.e., diagnostic learning), which indicates that they are sensitive to the directionality of the causal arrow. For example, Waldmann and Holyoak (1992) showed that blocking only occurs when the cues represent causes but not effects (see also Waldmann, 2000, 2001; Booth & Buehner, 2007; López, Cobos, & Caño, 2005, for more recent converging evidence). Causal directionality is an aspect of causation that cannot be reduced to covariation but is an integral component of causal model theory (see section on *Reasoning with Causal Models*, for more details).

Other demonstrations of the irreducibility of causation to covariation include the distinction between causal and non-causal (i.e., spurious) covariations (Cheng, 1997; Waldmann & Hagmayer, 2005), the distinction between covariation and causal power (Cheng, 1997), or the capacity of humans to derive differential predictions for hypothetical observations and interventions from identical covariation information (Gopnik, Glymour, Sobel, Schulz, Kushnir, & Danks, 2004; Sloman & Lagnado, 2005; Waldmann & Hagmayer, 2005). All these findings, which will be discussed later in greater detail, demonstrate how humans go beyond the information given, and infer causal properties on the basis of covariational learning input (see also Buehner & Cheng, 2005; Gopnik & Schulz, 2007; Waldmann, Hagmayer, & Blaisdell, 2006, for overviews).

(2) *Logical Theories*

A second class of theories that attempts to reduce causal reasoning to a domain-general theory are logical theories which model causal reasoning as a special case of deductive reasoning. Modeling causal reasoning in terms of propositional logic has proven problematic. From "If Cause A then Effect B" and "Cause A" we cannot infer Effect B ("modus ponens") in many real world cases because there may be additional

disablers of Effect B, or Cause A may be probabilistic, and therefore neither necessary nor sufficient (Cummins, 1995; Markovits & Potvin, 2001; Neys, Shaeken & Ydewalle, 2002, 2003; Quinn & Markovits, 1998). For the same reasons we cannot infer from the absence of Effect B to the absence of Cause A ("modus tollens").

Moreover, conditionals do not distinguish between causes and effect, and therefore can equally express "(1) If Cause A then Effect", and "(2) If Effect then Cause B", the latter type of rule being used often in diagnostic expert systems. However, the insensitivity of logical rules to causal directionality creates problems. For example, these two premises invite the transitive inference from Cause A to Effect, and then from Effect to Cause B, which is clearly an invalid inference since two causes of the same effect typically compete (Pearl, 1988). Although some of these problems may be solved if additional premises are added from background knowledge (Cummins, 1995), it can be doubted that this route fixes the underlying problem of neglecting causality (see Pearl, 1988, 2000).

A prominent example of a logical theory of causal reasoning is Goldvarg and Johnson-Laird's (2001) *mental model theory* of causation. The focus of this theory is reasoning with causal propositions. According to mental model theory people represent propositions by constructing mental models in which each model represents a possible state of affairs consistent with the premises. To distinguish between causes and effects, a temporal priority assumption is added according to which causes precede effects in time. For example, the full representation of proposition "A causes B" assumes three models in which (1) A precedes and co-occurs with B (i.e., a b), (2) the absence of A co-occurs with the absence of B (i.e., ~a ~b), and (3) the absence of A co-occurs with the presence of B (i.e., ~a b). The third possibility describes possible cases in which the effect is caused by some other factor.

Mental model theory assumes that people represent causality as deterministic. “A causes B” excludes as the only case the co-occurrence of A and non-B, thus assuming that A is sufficient (but not necessary) for its effect. According to the theory, “A causes B” can be distinguished from “A enables B” by modifying the third model (i.e., a ~b). Thus, enablers are necessary but not sufficient. Finally, preventing is modeled as causing the absence of the effect.

People often reason with reduced representations which only contain what is mentioned in the proposition (e.g., “A causes B”), which in the example would be the first model (a b). The other models are either neglected or there is a mental footnote referring to them. This assumption predicts that people often represent causal relations as the co-occurrence of two events and forget about other cases consistent with the causal claim.

Mental model theory is strictly Humean by reducing causality to temporal priority, co-occurrence, and determinism. Notably “A causes B” is represented by mental model theory the same way as “If A then B”. Thus, the theory clearly attempts to reduce causation to non-causal domain-general representations, and therefore shares many of the problems of other non-causal theories. For example, the theory cannot differentiate between causal and non-causal relations in which temporal priority holds. For example, barometer readings precede and co-occur with states of the weather but are clearly not causally related. Moreover, the theory does not distinguish between causes that are observed and causes that are generated by means of interventions (see Sloman & Lagnado, 2005). A further shortcoming of the theory is that the relationship between deterministic relations and probabilistic data is not worked out.

(3) *Probabilistic Theories*

Whereas associative theories update associative strength based on trial-by-trial learning, probabilistic theories pick up covariation information from frequency data,

which can be presented in various formats. Probabilistic theories typically assume that there are causes and effects, and that covariations need to be assessed in the cause-effect direction. Thus, they also represent a step in the direction of causal theories. However, they are still limited as a causal theory. The distinction between causes and effects is assumed, not an intrinsic part of a causal theory; the principal goal of research is to investigate which covariation metric people use to assess the strength of cause-effect relations. Thus again, causal relations are largely reduced to statistical covariations between causes and effects.

This research was pioneered by Kelley (1973) who postulated that people compute intuitive ANOVA analyses to make causal inferences. His theory proved very influential in social and cognitive psychology, and was formalized in various directions. A popular rule for measuring causal strength for a single cause-effect relation was the Δ*P* rule, which measures contingency as the difference between the conditional probabilities of the effect in the presence (*P*(e|c)) minus the absence of the cause (*P*(e|~c)) (Jenkins & Ward, 1965):

$$\Delta P = P(e|c) - P(e|\sim c) \qquad \text{(Equation 1)}$$

Thus, according to this rule, causes are *difference makers*, which raise (generative cause) or reduce (preventive cause) the probability of the effect. For example, eating nuts may be viewed as a cause of an allergy in a specific person if the person has a higher probability of having an allergy after having eaten nuts in comparison to not having eaten nuts. Interestingly, it can be shown that under certain conditions, the asymptotic strength computed by the Rescorla-Wagner (1972) model of associative learning is equivalent to Δ*P* (Danks, 2003).

The Δ*P* rule applies to single cause-effect relations, but often fails when multiple causes are present, which may introduce confoundings. For example, the ΔP rule does not predict the blocking effect that was mentioned above because it does not take into

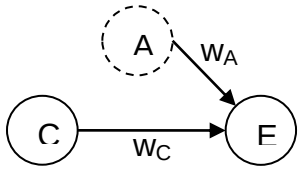
account competing causes. Therefore, Cheng and Novick (1992) have developed and tested a formal theory that computes probabilistic contrasts that take into account main effects and interactions, and therefore can deal with situations in which multiple causes converge on a joint effect.

The Δ*P* rule is not the only statistical rule that has been postulated as a measure of causal strength, there are numerous alternatives (see Perales & Shanks, 2007; Hattori & Oaksford, 2007, for overviews). However, the Δ*P* rule shares with most of these rules the problem of reducing causation to covariation between causes and effects.

From Covariation to Causal Power

The fact that covariation does not necessarily reflect causation can be seen in various cases. For example, barometer readings covary (to some extent) with the future weather, although there is no causal relation between the two variables. It is also possible that we observe zero covariations despite underlying causal power. If, for example, only people who do not suffer from headaches take aspirin, then aspirin will not covary with the amount of headaches, although they are both causally related. In this population aspirin simply cannot possibly display its potential power.

Figure 1. Parameterized causal model of a single causal relation. Nodes indicate the causal variables under consideration (*C* and *E*); the dashed node (*A*) represents additional unobserved causal factors influencing the effect; arrows indicate causal mechanisms having causal powers *w*.



(1) *Power PC theory*

In a seminal paper, Cheng (1997) has proposed a formal theory of how causal power can be estimated from covariation data when specific preconditions hold. The view that causality can be reduced to some metric of covariation was abandoned, and replaced by

the theory that causal power is a theoretical concept, which can be estimated under specific circumstances using covariation information and background knowledge. According to Cheng's (1997) power PC theory, causal power represents the probability of a cause, acting alone, to generate or prevent a target effect. Given that causes are never observed alone, this is a theoretical entity that needs to be inferred on the basis of observed data. To estimate causal power, several assumptions need to be made. Power PC theory partitions all (observed and unobserved) causes of effect *e* into the candidate cause in question, *c*, and *a*, a composite of all alternative causes of *e* (see Fig. 54.1). The term *a* represents the compound of all alternative enabling and generative causes. According to power PC theory people enter the task with the default assumption that *c* and *a* independently influence *e*, *a* produces but does not prevent *e*, that causal powers are independent of the frequencies of *c* and *a*, and that *e* does not occur without being caused. These default assumptions may of course be revised on the basis of contradicting evidence.

The unobservable probability with which *c produces e* is called *generative* power, represented by w_c. The generative power of the unobservable set of causes *a* is analogously termed w_a. On the condition that *c* and *a* influence *e* independently, it follows that

$$P(e|c) = w_c + P(a|c) \cdot w_a - w_c \cdot P(a|c) \cdot w_a \qquad \text{(Equation 2), and}$$

$$P(e|\sim c) = P(a|\sim c) \cdot w_a \qquad \text{(Equation 3).}$$

Equation 2 implies that effect *e* is either caused with a specific probability by *c* or by the unobservable *a* minus the overlap of the two event classes. The assumption embodied in Equation 2 that *c* and *a* are independent causes of *e* is an instantiation of a *noisy-OR* gate (see Cheng, 1997; Pearl, 1988; Griffiths & Tenenbaum, 2005). According to the noisy-OR integration rule, each cause has an individual chance to produce the effect, and the

causes do not interact when they occur simultaneously. The difference between $P(e|c)$ and $P(e|\sim c)$ can be abbreviated as ΔP (see Equation 1). Thus,

$$\Delta P = w_c + P(a|c) \cdot w_a - w_c \cdot P(a|c) \cdot w_a - P(a|\sim c) \cdot w_a \qquad \text{(Equation 4).}$$

Equation 4 shows why covariations do not directly reflect causality. 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. Ideally reasoners should restrict causal inference to situations in which *c* and *a* occur independently; that is, there is *no confounding*. In this special case, Equation 4 reduces to Equation 5:

$$w_c = \Delta P/(1- P(e|\sim c)) \qquad \text{(Equation 5)}$$

Of course, since *a* is unobserved, reasoners cannot be sure whether *a* and *c* are indeed independent in a particular case. Cheng (1997) assumes that independence is the *default* assumption people make when there is no evidence to the contrary (but see Luhmann & Ahn, 2007). A strategy to ensure independence is to manipulate the presence or absence of *c* by means of an intervention, as scientists do in experiments. These manipulations make sure that *c* and *a* occur independently even when *a* is unobserved.

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

Confounding may of course be due to *observed* alternative causes as well, not only to unobserved ones. Research on confounding by observed causes has shown that people are often aware of the confounding and therefore tend to create independence by holding the alternative cause constant, preferably in its absent value. For example, Waldmann and Hagmayer (2001) have shown that when assessing causal strength between a target cause and a target effect learners hold a third event constant only when it

is an alternative cause but not when it is causally irrelevant or a causal effect (see also Goedert, Harsch, & Spellman, 2005; Spellman, 1996).

Numerous studies have tested power PC theory and related accounts. The typical research strategy in this field is to present participants with various causal scenarios in which the contingencies between a target cause and an effect are varied by manipulating the conditional probabilities of the effect in the presence versus absence of the cause. After the learning phase causal strength estimates are requested by the participants. To discriminate between the competing theories, contingencies are chosen that entail different causal strength estimates in the competing theories. These studies have generally shown that in many circumstances learners indeed try to estimate causal power and take into account alternative causes when presented with covariation data (Buehner, Cheng & Clifford, 2003; Cheng, Novick, Liljeholm, & Ford, 2007; Hagmayer & Waldmann, 2007; Novick & Cheng, 2004; Wu & Cheng, 1999), although there are also situations in which learners seem to be trying to estimate covariation or other statistics (see Cheng & Novick, 2005; Hattori & Oaksford, 2007; López & Shanks, 2008; Luhmann & Ahn, 2007; White, 2003).

In sum, Cheng's (1997) theory computes point estimates of w_c, causal power. Formally the theory provides an answer to the query for the maximum likelihood estimate of w_c. However, power PC theory does not take into account the *uncertainty* of inductive causal inference. In particular, the power PC estimate is insensitive to sample size, and other sample statistics that should affect inductive inference.

(2) *Causal Support Theory*

Griffiths and Tenenbaum (2005) analyzed causal inference in the context of their *causal support model* which takes into account the uncertainty of parameter estimates by considering distributions of parameters in contrast to point estimates (see also Griffiths, Kemp, & Tenenbaum, 2008). Like Cheng's (1997) theory this account is derived from

normative considerations about rational inference; both theories postulate that everyday learners strive to reason rationally if possible. Thus, the primary goal of these theories is to provide a computational account of human reasoning, not a theory of cognitive mechanisms.

Whereas Cheng's (1997) focus were judgments of causal strength, which were modeled as parameter estimation tasks, the causal support model focuses on the assessment of the likelihood of the presence of a causal link between a target cause *c* and an effect *e*. Griffiths and Tenenbaum argued that this is actually the question learners try to answer when estimating causal power. Within a Bayesian account structure judgments are a case of *model selection* (see Mackay, 2003). The causal support model aims to contrast the evidence in favor of a causal model in which a link exists between *c* and *e* (Model 1; see Fig. 54.1) with a causal model in which these two variables are independent, and *e* is only influenced by the background causes *a* (Model 0). As in power PC theory, it is assumed that there are additional hidden causes *a* in the background which affect the base rate of the effect. Also it is assumed that *c* and *a* are, in the case of generative causation, combined by the noisy-OR gate.

The center of Bayesian inference is Bayes' rule,

$$P(H \mid D) = \frac{P(D \mid H)P(H)}{P(D)} \qquad \text{(Equation 6)}$$

where *H* denotes a hypothesis, and *D* denotes observed data. Bayes' rule provides a formal tool for modeling the inference concerning the posterior probability of a hypothesis, $P(H \mid D)$, with $P(H)$ representing the prior belief in the hypothesis and $P(D \mid H)$ representing the likelihood of the data given the hypothesis.

In general, there are several competing hypotheses that may explain a given set of data, like Model 1 and Model 0 for a simple causal relation. To compute how likely a model is in comparison to other models, a ratio representation of Bayes' rule is used. In

causal support theory the decision for a model is based on the posterior probability ratio of Model 1 and Model 0 by applying Bayes rule (cf. Equation 6):

$$\log \frac{P(Model1 \mid D)}{P(Model0 \mid D)} = \log \frac{P(D \mid Model1)}{P(D \mid Model0)} + \log \frac{P(Model1)}{P(Model0)} \ . \qquad \text{(Equation 7)}$$

Thus, the log ratio of the posterior probabilities (left side of Eq. 6) equals the sum of the log ratio of the likelihoods of the data given each model (first expression on right side) and the log ratio of the models' prior probabilities (second expression). Assuming equal base rates of Models 1 and 0, causal support is determined by the log likelihood ratio,

$$\text{support} = \log \frac{P(D \mid Model1)}{P(D \mid Model0)}. \qquad \text{(Equation 8)}$$

Thus, support represents a measure of the degree of evidence that data *D* provide in favor of Model 1 over Model 0. The computations of the likelihoods of the data *D* given the models take into account the uncertainty attached to parameter estimation. The likelihood of *D* given Model 1 versus Model 0 are computed by averaging over the unknown parameter values, causal strengths w_a and w_c, which can vary between 0 and 1.

Griffiths and Tenenbaum (2005) presented experiments providing evidence for their causal support theory, generally using the same research paradigm used to test power PC theory: Participants are confronted with contingencies which entail different patterns of strength estimates in the competing theories (e.g., causal support theory vs. power PC theory). These experiments showed sensitivity to sample size, and demonstrated good fits between power estimates and the Bayesian causal support measure. Thus, these studies suggest that people indeed take into account the uncertainty of statistical inference in causal induction.

Lu, Yuille, Liljeholm, Cheng, and Holyoak (2008) have further developed Bayesian models of power estimation, considering different variants of priors for the causal models and their parameters. Various experiments showed that models incorporating a simplicity

bias along with the assumption that causes tend to be strong ("strong and sparse bias") provided the best fit to the data.

(3) *Summary*

In sum, the literature on causal power estimation suggests that people often try to go beyond covariation information and are sensitive to causal power. However, there are different ways a learner may interpret a power query. For example, she may provide an estimate of the strength parameter, assess the likelihood of a causal link, or interpret the test question as a request to judge the extent to which the observed cause actually is responsible for the occurrence of the effect in the case at hand (see Cheng & Novick, 2005). The complexity of the task also affects the inference strategies, which sometimes prevents reasoners from reaching their goals to estimate power (see De Houwer & Beckers, 2003; Waldmann & Walker, 2005). Despite the wealth of studies on causal strength estimations, little is known about the interplay between inference goals and task factors in determining how people estimate causal strength.

Causal Mechanisms, Processes, and Forces

The theories we have discussed so far are all variants of theories that view causes as difference makers. In probabilistic theories, causes change the probability of the effect, counterfactual and logical theories compare the hypothetical or actual absence of the cause with its presence (Dowe, 2000). These accounts can be contrasted with mechanistic *process theories*, which focus on processes and mechanisms initiated by causal events. A popular recent philosophical example of such a theory was developed by Dowe (2000), who characterized causal processes in terms of the transmission of a *conserved quantity*, such as linear momentum, charge, and so on. This theory can be applied to physical processes but it is less clear how such an account would model other domains (e.g., economy). Moreover, it seems unlikely that people who are not scientists know much about conserved quantities. A more general account, which has also been adopted in

psychology are theories that focus on the *mechanisms* relating causes and effects (e.g., Machamer, Darden, & Craver, 2000).

(1) *Mechanisms and Covariation*

Although Rozenblit and Keil (2002) have shown that people have little knowledge about the mechanisms underlying artifacts from everyday life, it still seems plausible that mechanism information is considered important when it is available. Most of us certainly believe that causes are connected with effects by some mechanism, even when we do not really care to get to know the details.

A number of studies have investigated how covariation and mechanism information interact. Knowledge about causal mechanism typically increases estimates of perceived correlations relative to the objective covariations in the data (Koslowski, 1996) – even when the covariation is zero (Chapman & Chapman, 1967, 1969). Ahn, Kalish, Medin, and Gelman (1995) employed an information search paradigm to pit mechanism against covariation information. In their experiments participants were confronted with various positive and negative fictitious facts, such as the promotion of Mary last year. Participants were requested to ask further questions which would help them to find out which causal factor is responsible for the target event. The general finding of this study was that participants were more interested in potential mechanisms rather than covariation. For example, participants were more interested in whether Mary did something special that would motivate her raise, rather than information about how many of her colleagues also got raises.

Fugelsang and Thompson (2003) proposed a dual-process theory which claims that causal judgments are influenced by two independent sources, covariation information and mechanism knowledge. The results were not entirely conclusive (see Perales, Shanks, & Lagnado, 2010), but the authors interpret their findings as showing that new data is weighted more heavily in causality judgments when there is a plausible mechanism than

when the mechanism is implausible. In contrast, new covariation data is simply combined with information about covariation in the past, regardless of whether the new and old covariation match or mismatch.

All these studies show that people care for mechanism information. However, pitting mechanisms against covariation does not reflect the current state of the field anymore. The previous sections have shown that modern theories of probabilistic causation do not claim anymore that causation can be reduced to covariation. Covariation rather is an empirical indicator of causal relations, which is especially important in situations in which we do not have prior knowledge and need to induce causal knowledge based on data (see Cheng, 1993). This view is certainly consistent with the findings that knowledge about mechanisms, which might have been induced in previous learning contexts, may influence current judgments. It also makes sense to weigh well established knowledge more than current data, which is typically noisy.

Moreover, covariation assessment itself is intrinsically linked to mechanism knowledge. If people observe a continuous sequence of events, the number of possible covariations that could be computed clearly surpasses their information processing capacity. Mechanism knowledge can place constraints on the events we consider for covariation assessments. If we suddenly experience nausea we may hypothesize a drug we ingested two hours ago as the cause but not a food item we ate one minute ago, or various other irrelevant events. A number of studies have shown that people use intuitions about the temporal delays of different mechanisms when making covariation assessments (Buehner & May, 2002, 2003; Buehner, 2005; Greville & Buehner, 2007, 2010; Hagmayer & Waldmann, 2002).

Another reason why mechanism and covariation theories need not be seen as competitors anymore, is that the focus on single cause-effect relations has been replaced by a greater interest in other causal models, such as causal chains (see section on

Reasoning with Causal Models). Within probabilistic theories, mechanisms can be modeled as causal chains in which multiple events form a sequence (see Fig. 54.2). When information about the potential alternative causal explanations is already available, it seems reasonable to ask for information about the chaining of causal relations, rather than for contrast information that is already known. For example, in the study of Ahn et al. (1995) participants probably already knew possible causes for salary raises, and therefore needed information which restrict the possible explanations.

(2) Causal Forces

So far we have discussed theories that postulate causal relations between classes of events, for example between smoking and lung disease. However, we can also ask about the actual cause in specific situations (i.e., singular causation), for example whether the lung disease Peter contracted last year was due to his smoking, exposure to asbestos, or some other cause. According to probabilistic theories, both levels are clearly related, we need generic level knowledge to answer questions about actual causes but nevertheless both levels need to be separated. Thus, although in Peter's case we only have observed the presence of smoking and lung disease, which by itself does not ensure a causal relation, smoking is a candidate for his lung disease because we know that in the reference class to which Peter belongs smoking and lung disease are causally related (see Pearl, 2000).

However, there is an alternative view: Some psychologists have argued that the specific level in which no covariation information is available is the primary level, and that we derive generic level conclusions from collections of specific, singular cases. An example of such an approach are force theories which assume that people represent singular causal events as generated by hidden forces. Whereas philosophical theories of mechanisms and processes try to model causation in terms of normative scientific

theories, the forces postulated by the psychological theories bear more similarity to medieval impetus theories (McCloskey, 1983) than to modern Newtonian physics.

White (2006, 2009) uses Michotte type launching events to demonstrate the difference between intuitive causal representations and physics (see also White, 2005). In Michotte's (1963) famous demonstrations of phenomenal causality participants observed moving objects. For example, in a launching scenario, Object A moves towards Object B, and touches it. This stops Object A and sets Object B into motion at the same or a slightly lesser speed. Observers typically describe this scenario as a case in which the movement of Object B is caused by Object A (i.e., launching). Although according to Newtonian physics the force on body B exerted by body A is equal in magnitude but opposite in direction to that on body A exerted by body B, observers often see Object A as the cause and Object B as the effect (causal asymmetry). Nobody would describe the scenario as a case of Object B stopping Object A, although this would be a legitimate description.

The impression of causal asymmetry is also reflected in judgments about force. White (2009) presented participants with different launching events and asked them provide estimates of the relevant underlying forces. The results showed that in such events more force is attributed to Object A than Object B, and that Object A is viewed as active and exerting a *force* on Object B, whereas the initially stationary Object B is viewed as inactive, exerting *resistance* to being moved. Thus, causal interactions are perceived as the result of the opposition between forces of agents (e.g., Object A) and resistance of patients (e.g., Object B). Within White's theory, force and resistance are theoretical concepts that need to be estimated based on observable data. White (2009) shows that different observable kinematic features, such as the velocity of objects before and after contact, are the basis for force and resistance attributions.

How can the impression of causal asymmetry be explained? In developmental psychology the hypothesis is popular that force attributions in launching scenarios may

be due to an innate module specialized for causal analysis (see Carey, 2009; Leslie & Keeble, 1987). White (2009) disagrees and proposes the theory that our haptically experienced actions are the primary sources of intuitions about force and resistance. When we manipulate object we experience the required force, and take into account properties of the objects, which are experienced as resistance. These sensorimotoric schemas are used to make sense of analogous causal interactions which we only passively observe.

White's (2009) theory is restricted to launching and similar events. A more general theory that aims at elucidating our understanding of abstract causal concepts, such as *cause, prevent,* and *enable,* is Wolff's (2007) theory of *force dynamics* (see also Wolff & Song, 2003; Talmy, 1988). Wolff (2007) also focuses on specific causal events, but studies a wide range of scenarios from different domains. As in the theories of White (2009) and Talmy (1988), two entities are distinguished, which Wolff calls affectors and patients (i.e., the entity acted upon by the affector). Force theory states that people evaluate configurations of forces attached to affectors and patients, which may vary in direction and degree, with respect to an endstate, that is, the possible result. Forces can be physical, psychological (e.g., intentions) or social (e.g., peer pressure). Causal relations are analyzed in terms of three components, (a) the tendency of a patient for an endstate, (b) the presence or absence of concordance between affector and patient, and (c) the degree to which the endstate is reached. Table 1 summarizes the predictions for how people use the concepts *cause*, *prevent*, and *allow (enable)*.

Table 1: Force dynamic analysis of the meaning of causal concepts

	Patient tendency for endstate	Affector-patient concordance	Endstate approached
Cause	No	No	Yes
Allow (enable)	Yes	Yes	Yes
Prevent	Yes	No	No

For example, force theory would represent the singular causal fact "Winds caused the boat to heel" in terms of a patient (the boat) that had no tendency to heel (Tendency = No), the affector (the wind) acted against the patient (Concordance = No), and the result (heeling) occurred (Endstate approached = Yes). In contrast, "Vitamin B allowed the body to digest" describes a scenario in which the patient (body) already has a tendency toward the endstate (digest), which is reached in accordance with the affector (vitamin B). Preventing, in contrast, refers to cases in which an affector exerts a force that counteracts the tendency of a patient toward an endstate (e.g., "Wind prevented the boat from reaching the harbor").

Empirical support for the model was provided in a series of experiments in which participants made judgments about 3-D animations of realistically rendered objects (e.g., moving boats on a lake) with trajectories that were wholly determined by the force vectors entered into a physics simulator. For example, participants were asked to judge whether they are viewing a case of *cause*, *enable*, or *prevent* (see Wolff, Barbey, & Hausknecht, 2010; Lombrozo, 2010, for further developments).

Force theories represent an important novel class of theories which highlight aspects of causality that have largely been neglected in previous theories. Causes have mostly been viewed as entities that make a difference with respect to their effects, but little is said about how they make a difference, and why a specific functional form is observed between causes and effects. Force theories also point to the importance of empirical indicators of causation beyond (or instead) of covariation (e.g., velocity). However, research on force theories is still in the beginning stage so that a number of questions remain unanswered. For example, it is unclear how general these theories are. Wolff (2007) thinks that force theories can replace other theories, but at this point it is not clear whether they can successfully model all kinds of causal relation. The model is very convincing in cases of perceptions of cases of singular causation (e.g., moving objects) in

which attributions of forces and endstates seem natural, but it is less clear whether all causal relations, especially generic ones, are also represented this way. Do we represent causal claims such as "social status influences success" or "interest rates influence spending" in terms of forces and resistance, or forces and endstates? Although this may be the case, there is no direct empirical evidence for such a claim. Sloman, Barbey, and Hotaling (2009) have proposed a successful causal model theory of cause, enable, and prevent that does not contain references to forces and tendencies (see also Cheng & Novick, 1991).

Wolff's (2007) version of force theory is more general than White's (2009) analysis of launching events, which comes at the cost that it is unclear how the different theoretical concepts are estimated. White has elaborated such an account for launching events but kinematic information (e.g., velocity) is obviously only relevant in a restricted set of domains. Other empirical indicators need to be discovered for other manifestations of causal forces.

Force theories have tried to separate themselves from theories that use covariation as input to causal power assessments. However, it seems implausible to generally dismiss this information. For example, it seems unlikely that people would make force attributions when considering the claim "Vitamin B allowed the body to digest" (Wolff, 2007), when there is no prior covariational knowledge that supports vitamins as potential causal agents. Covariation information allows people to infer causal relations between variables even when no cues relevant for phenomenal causality are present. Schlottmann and Shanks (1992), for example, showed that people can use correlations between color change cues and movements to judge an underlying causal relation although color changes did not lead to the perception of phenomenal causality.

Reasoning with Causal Models

Force and mechanism theories have so far primarily been applied to singular causal relations, and do not address the question how intuitions about forces and mechanisms can be related to covariation information. Causal model representations provide tools to integrate basic intuitions about causal relations with inference and learning methods. A further advantage of causal models is that they can be applied to complex networks of causes and effects, thus overcoming the restrictive focus on individual causal relations.

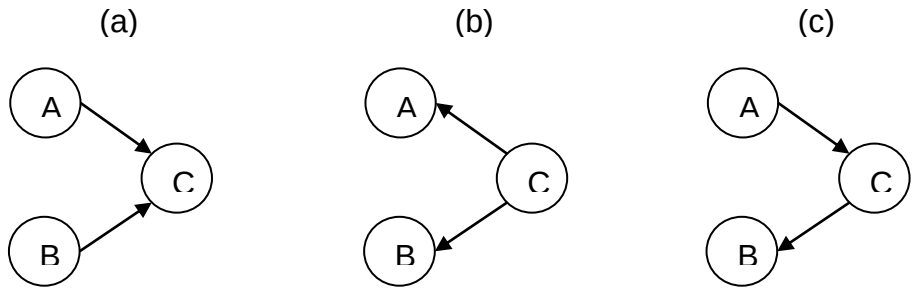
Networks of cause-effect relations lend themselves to graphical representations, which have been introduced in philosophy (Reichenbach, 1956), and further developed to Bayes net or causal Bayes net theory in both philosophy and Artificial Intelligence research (e.g., Pearl, 1988, 2000; Spirtes, Glymour, & Scheines, 1993; see Koller & Friedman, 2009, for a recent textbook). Although Bayes nets have primarily been developed as a practical tool for automated inference and data mining, they have also inspired psychological research (see Gopnik et al., 2004; Sloman, 2005, for overviews). Causal Bayes net theory is not a unified theory but integrates various more specific tools, which neither are nor are meant to be plausible as psychological theories. Moreover, a number of formal assumptions are controversial as psychological claims. Therefore, we will use the more neutral term *causal model* in the following discussion except when discussing the specific assumptions underlying causal Bayes nets.

Figure 54.2 shows examples of causal models with three events. In general, causal models include variables that represent causes and effects along with arrows representing direct causal relations. These graphical representations of the structure of the causal models provide useful information about conditional dependence and independence relations. Assuming the *Markov* condition, which states that each variable, conditional upon its direct causes, is independent of all other variables except its direct and indirect effects, the common cause model in Figure 54.2b implies, for example, that effects A and B of the common cause C are correlated but become independent conditional upon their

cause. Similarly, in the causal chain model (Fig. 54.2c) the final effect B is correlated with the initial cause A and the intermediate cause C but becomes independent from the initial cause when the intermediate cause is kept constant. Finally, the common effect model (Fig. 54.2a) implies that the causes A and B are marginally independent from each other but become dependent once the state of the effect C is known. This feature of common effect models explains the familiar intuition of *explaining away* or *discounting*. For example, once we know that we have nausea due to contaminated food, other potential causes, such as an infection, become less probable.

Apart from structure information, causal models also contain parameters that can be estimated based on learning data. These parameters (e.g., w_c in Fig. 54.1) quantify estimates of causal power of generative or preventive causes, and assumptions about the way multiple causes are integrated (e.g., linear additive integration by noisy-OR). Causal models can be learned, and they can support predictive inferences from causes to effects, or diagnostic inferences from effects to causes. The following sections will give an overview of the most important psychological findings motivated by causal model theories.

Figure 2. Basic causal models with three causal variables, (a) common effect, (b) common cause, (c) causal chain model.



(1) Causal arrows

One of the most fundamental properties setting causal relations apart from mere covariations is the directionality of the causal arrow. Regardless of the order in which causal events are experienced, causal relations are directed from causes to their effects. Causes generate their effects but not vice versa, which can be explained by the unidirectionality of the underlying mechanism.

Waldmann and Holyoak (1992; Waldmann, Holyoak, & Fratianne, 1995) introduced *causal model theory* in cognitive psychology, hypothesizing that learners use abstract causal knowledge about causal networks guiding their processing of the learning input. The general idea of their experimental paradigm was to present participants in different learning conditions with identical covarying events but manipulate the intuitions about which events represent causes and which effects (see also the section on *Associative Theories*). For example, Waldmann (2000, 2001) presented learners first with cues that represented substances in hypothetical patients' blood and then gave feedback about fictitious blood diseases. Two conditions manipulated—through initial instructions—whether learners interpreted the substances as effects of the diseases (common cause model) or as causes (common effect model). Thus, learners represented identical cues either as causes which they used to predict a common effect (i.e., predictive learning) or as effects to diagnose a common cause (i.e., diagnostic learning). In the test phase learners were asked to estimate the strength of the relation between substances and disease. The results showed that causal models guided how the learning input was processed. Learners treated the substances as potentially competing explanations of the disease in the common effect condition, whereas the substances were treated as collateral correlated evidence of a common cause in the contrasting condition. Consequently, effects conforming to patterns entailed by cue competition (e.g., blocking) resulted only when the cues represented causes but not effects. These findings demonstrate that people

do not simply associate cues with outcomes but represent the learning events within causal model structures.

Further evidence for sensitivity to the direction of the causal arrow comes from a *semantic memory* study by Fenker, Waldmann, and Holyoak (2005). Questions referring to the existence of a causal relation between events described by a pair of words (e.g., spark-fire) were answered faster when the first word of a pair of words referred to a cause and the second word to its effect than vice versa (i.e., fire-spark). No such asymmetry was observed, however, with questions referring to the associative relation between the two words.

Evidence for the asymmetry of causes and effects can also be found in research on *inductive reasoning* about properties. For example, undergraduates are more likely to infer that "lions have enzyme X" from the premise "gazelles have enzyme X" than vice versa because their knowledge about food chains makes it easier for them to imagine the transmission from gazelles to lions than the other way around (Medin, Coley, Storms, & Hayes, 2003).

Finally, Ahn and colleagues, using category membership judgments, have demonstrated that causes often carry more weight than effects in causal *categories* (causal status effect). For example, in Ahn, Kim, Lassaline, and Dennis (2000) participants were instructed about artificial animal categories that were described as possessing three features: eats fruit (X), has sticky feet (Y), and builds nests on trees (Z), which are causally linked within a chain. When presented with items in which one feature was missing subjects showed that they found the item least likely to be a member of the category when X was missing and most likely when only Z was missing (see also Kim & Ahn, 2002, for examples from clinical psychology). The theoretical source and generality of these findings is currently under debate. Rehder and Kim (2006) have questioned the generality of the causal status effect, and have shown that a common effect of multiple

alternative causes may receive more weight than either of its causes. At any rate this set of findings does provide additional evidence for the psychological difference between cause and effect representations.

(1) Causal Structure

One of the main strengths of causal model theories is that they do not only focus on models in which one or more causes converge on a common effect, but also on more complex causal models. One key advantage of causal model representation is their parsimony. Only direct causal relations are represented (by arrows), whereas covariations between indirectly linked events can be computed from the model based on information about the strengths of the direct relations and the structure of the model. A number of studies have investigated whether people are capable of deriving predictions for indirect relations.

For example, Waldmann et al. (1995) have used a *learning* paradigm in which participants were presented with multiple cues which were either described as causes of an effect (common effect model), or as effects of a single cause (common cause model). For example, in one experiment the task was to learn to classify stones as magnetic or non-magnetic based on the spatial orientation of surrounding iron compounds. In the common effect condition the iron compounds were described as potential causes of the stones being magnetic, whereas in the common cause condition the magnets were characterized as influencing the spatial orientation of the surrounding iron compounds. Learning difficulty was assessed by measuring the mean number of errors until a criterion was reached. Across different conditions, it was manipulated whether the cues were correlated or independent. The findings showed that participants in the common cause condition learned faster to predict the outcome when the cues were correlated, whereas independent cues yielded faster learning with the common effect model. This finding is consistent with the fact that common cause models entail correlations between their

effects, whereas the default assumption of common effect models is that the causes are independent.

Rehder and colleagues have presented a number of studies showing sensitivity to structural implications of causal models in *categorization* tasks. For example, Rehder (2003a, b) presented subjects with fictitious categories of stars (e.g., Myastars) which had five different binary features (e.g., ionized helium, very hot, high density, etc.), which varied between abnormal and normal values (see also Rehder & Hastie, 2001). Additionally instructions about the causal model connecting these features were provided (common cause model; chain; common effect model). Then different exemplars with different feature configurations were presented along with the task to rate the degree of category membership (i.e., typicality). One important finding was that subjects rated exemplars more typical that were consistent with the structural implications of the instructed causal model (coherence effect). For example, subjects expected that in causal chains or common cause models it is very likely that either all features are normal or all abnormal. Thus, subjects expected correlations between indirectly linked events, which is in line with the structural implications of causal models (see also Rehder & Kim, 2006, for more complex causal models).

Sensitivity to the structure of the underlying causal model has also been demonstrated in *inductive inference* tasks (see Rehder & Hastie, 2001; Rehder, 2009). Rehder (2009) used his paradigm in which subjects were informed about fictitious categories with multiple features, for example, Romanian Rogos (a type of automobile). After learning about Rogos, participants in one of the experiments were presented with a series of trials in which they were told about one novel feature, for example a zinc laden tank. The crucial manipulation involved the causal role of this new feature. Some participants were told that zinc laden tanks caused one of the familiar features of Rogos, whereas others were instructed that these tanks were effects of one of the features. When subjects were

asked how prevalent the new feature was within the category of Rogos they gave higher ratings when it was causally linked to a common feature than to a rare feature regardless of its causal status, which is predicted by causal model theories.

(2) Observing versus Intervening

One of the key differences between causal models and probabilistic or associative models is that they support inferences about the consequences of actions. Probabilistic or associative models tell us how variables are correlated but they do not distinguish between spurious non-causal and causal correlations. Interventions in a cause lead to its effects, but this effect cannot be accomplished by an intervention in a spurious correlate. For example, barometers are spuriously correlated with weather, and therefore can be used to predict the weather. However, tampering with the barometer does not change the weather because they are not causes of the weather. Thus, causal representations are crucial for correct predictions of the consequences of actions. Another feature of interventions, which are captured by causal Bayes net theories, is the fact that interventions, which deterministically and independently change the states of a target variable, remove all causal influences on the variable that is the target of the intervention. If we tamper with the barometer, then its reading is solely influenced by our manipulation but not by its usual cause, atmospheric pressure (Pearl, 2000; Spirtes et al., 1993; Woodward, 2003). Causal Bayes nets model interventions in variables as a removal of all causal arrows that normally influence this variable ("graph surgery")(see also Waldmann, Cheng, Hagmayer, & Blaisdell, 2008, for an explanation of interventions in terms of discounting and explaining away).

Do people distinguish between observations and interventions as causal Bayes nets predict? Research indicates that they do (see Hagmayer, Sloman, Lagnado, & Waldmann, 2007, for an overview). Sloman and Lagnado (2005) studied a number of causal models and interventions that removed individual events in *reasoning* tasks. For example,

participants were told that 'When A happens, it causes B most of the time; when B happens it causes C most of the time, A is present and C is present.' Then they were asked either to imagine that B was observed to be absent, or to imagine that B was actively prevented from occurring. In both cases participants were requested to draw inferences about A and C. Subjects generally predicted the absence of C but they predicted the absence of A only when the absence of B was observed but not when it was removed by an intervention ("undoing"). Thus, subjects had the intuition that the cause of B is unaffected if B is actively removed. Probabilistic theories or theories of propositional logical reasoning including mental model theory do not predict this finding.

Waldmann and Hagmayer (2005) showed that people also distinguish between observations and interventions when subjects receive both instructions about the structure of the underlying causal model and *learning* data about probabilistic relations between causal events (i.e., covariations). The study demonstrates that people use the learning data to infer the parameters of the underlying causal model, and are able to use the parameterized causal model to derive estimates about the probabilities resulting from hypothetical interventions and observations. These findings have been confirmed for more complicated models involving confounding causal pathways and a broader variety of learning procedures and species (Blaisdell, Sawa, Leising, & Waldmann, 2006; Leising et al., 2008; Meder et al., 2008, 2009a).

It is particularly important to distinguish between interventional and observational probabilities when we make *choices*. When we decide which action to take, it is the interventional probabilities that matter, and not simply the observed probabilities (Joyce, 1999). Hagmayer and Sloman (2009) examined the question whether subjects use interventional probabilities in decision making. For example, participants were told that men who do the chores are substantially more likely to be in good health than men who do not. In addition, participants were provided with causal explanations for this fact.

Participants in one group were informed that the relation is due to a common cause (degree of concern) that causes men to help at home and to care about their health. Another group was told that doing the chores is a form of exercise that positively affects health. Participants were sensitive to the causal structure underlying the probabilistic relation and preferred to start doing the chores when there was a direct causal link. Moreover, participants differentiated between an action that was merely observed versus an action that was actively chosen in their estimates of the probability of the desired outcome. Only choices were treated as hypothetical interventions within a causal Bayes net.

The difference between observations and interventions is not only important in reasoning and decision making, but also can aid learning. Interventions often allow us to discriminate between alternative causal models. For example, a correlation between bacteria and ulcer does not tell us whether bacteria cause ulcer or whether there is a common cause of both. An intervention (e.g., removing the bacteria) can tell us whether bacteria are the cause. A number of studies have shown that learners can use interventions to aid their learning (Gopnik et al., 2004; Lagnado & Sloman, 2004; Steyvers, Tenenbaum, Wagenmakers, & Blum, 2003).

(3) Knowledge-based causal induction

Many Bayes net researchers in computer science have focused on the development of statistical tools for scientific research which require minimal prior knowledge. An example of a learning method that uses minimal knowledge are *constraint-based algorithms* which allow us to induce causal structures on the basis of the patterns of statistical dependency within a set of variables (Pearl, 2000; Spirtes et al., 1993). A second approach to structure learning is framing the task in terms of Bayesian inference.

In this approach the learner must determine the likelihood of a structure hypothesis given data. To achieve generality, all possible structures along with minimal assumptions about the parameters are being considered in these statistical theories. Both approaches make few assumptions about the domain and therefore need vast amounts of reliable data to accomplish their goal.

In psychology there has been a debate whether these algorithms are plausible models of human learning (see Gopnik et al., 2004; Griffiths & Tenenbaum, 2009). The high data demands of the Bayes net algorithms cast doubt on the psychological plausibility of these models given that people often successfully acquire causal knowledge on the basis of very few learning trials. Empirical studies confronting subjects with the task to induce causal models based on covariation data alone have generally shown poor performance with observational data although these studies typically presented a small number of variables along with information about a limited number of alternative models to be considered (Steyvers et al., 2003). A number of studies have shown that allowing subjects to intervene helps (Gopnik et al., 2004; Steyvers et al., 2003). Kushnir, Gopnik, Lucas, and Schulz (2010) could even show that people can induce hidden causes when people received information about salient (deterministic) covariations. In sum, there is some evidence that people can use covariation information to induce causal structures but performance is poor unless learning conditions are favorable.

The implausibility of domain-general algorithms of structure induction has led Waldmann (1996) to propose the view that people generally use prior hypothetical knowledge about the structure of causal models to guide learning in a top down fashion ("knowledge-based causal induction")(see also Lagnado, Waldmann, Hagmayer, & Sloman, 2007). People use various cues which guide their initial hypotheses about causal structure, including temporal order, interventions, and prior knowledge. Temporal order is a potent cue to causal order (causes typically precede their effects). Prior knowledge

may, however, override the temporal cue. For example, a physician may see the symptom of herpes (i.e., effect information) before test results about the cause come in but can nevertheless form the correct causal hypothesis that a herpes virus has caused the observed skin symptoms and not vice versa. Numerous studies have shown that people can disentangle temporal from causal order (see Waldmann et al., 2006). Interventions are a further potent cue that allows us to distinguish causes from effects. Inducing causal structures based on cues simplifies the learning of causal strength parameters because learners may now focus on individual links rather than encoding patterns of dependency (see Fernbach & Sloman, 2009; Waldmann et al., 2008).

Assumptions about integration rules underlying multiple causes (i.e., functional form) also provide knowledge-based constraints on learning. Instead of considering large numbers of possible parameterizations as in Bayesian structure learning, the task may be simplified by considering only plausible default rules (see Griffiths & Tenenbaum, 2009). A typical default assumption underlying common effect models with generative causes is that the multiple causes independently influence their effect (noisy-OR schema)(see Cheng, 1997; Griffiths & Tenenbaum, 2005; Lu et al., 2008).

Although independent causal influences may be the default assumption, multiple causes can also interact (see Novick & Cheng, 2004). Interactions require more complex representations because the causal influence of a configuration of multiple causes cannot be reduced to a function of the individual causes. Empirical work on interactions and integration rules has only just begun. One factor influencing assumptions about integration rules is prior learning. Beckers et al. (2005) and Lucas and Griffiths (2010) have shown that people can transfer non-additive integration rules from a previous learning context to the present task (see also Shanks & Darby, 1998).

Another factor influencing the choice of integration rules is domain knowledge. Waldmann (1996, 2007) presented subjects with colored liquids which potentially

affected the heart rate when consumed. Different conditions showed that subjects tended to add the individual influences when the liquids were introduced as drugs with different *strengths*. However, subjects averaged the influences when they believed that the heart rate is dependent on the *taste* of the liquids. Here domain knowledge about the different properties of extensive and intensive physical quantities determined the integration rule.

Griffiths and Tenenbaum (2007, 2009) have proposed *hierarchical* Bayesian inference to model knowledge-based (or theory-based) causal induction. The basic idea is that probabilistic inference is carried out at multiple levels of abstraction which influence each other and are updated simultaneously. In causal learning these levels include the data, alternative causal models, and the theory level which encodes knowledge about the types of events (e.g., causes vs. effects), the plausibility of a causal relationship, and the functional form of these relationships (e.g., noisy-OR). In the hierarchical probabilistic model, the theory level at the top defines a probability distribution over causal model hypotheses, as each hypothesis defines a probability distribution over data. These different levels are then updated based on the data using the basic Bayes' rule (see Equation 5). Hierarchical Bayesian models speed up learning because the range of alternative hypotheses being considered is constrained by both the data and the theory level. Moreover, updating occurs on each level so that it is not necessary to assume fixed innate knowledge. Biases encoded on the theory level can be changed when the data disconfirms them. Finally, hierarchical models have the advantage of being able to generalize to new contexts. In a specific learning context both specific causal model knowledge and abstract theory knowledge are being updated. This abstract knowledge can in turn influence learning in novel domains, and thereby explain transfer effects (Lucas & Griffiths, 2010).

(4) Summary

Causal models provide a fruitful framework for modeling reasoning and learning in causal domains. They overcome the traditional restrictive focus on single cause-effect relations and have motivated studies on more complex causal scenarios. Moreover, the simultaneous development of causal model theory in both computer sciences and psychology has proven mutually fruitful, although it turned out that not all developments in engineering yield plausible psychological theories. Causal model theories go beyond traditional non-causal theories in that both the postulated representation and the inference and learning mechanisms are causally motivated. Causal models embody information about the structural difference between causes and effects, interventions and observations, and combine causal structure information with parameters reflecting causal power. Moreover, hierarchical knowledge-based theories allow for the integration of abstract theoretical knowledge with the processing of covariation information.

Unlike, for example, force theories, causal models provide an integrated computational mechanism which connects mechanism knowledge with covariation information. However, thus far only very basic information about mechanisms is expressed in causal models. The asymmetry of causal relations is encoded in arrows, which Pearl (2000) has interpreted as "mechanism placeholders." Prior knowledge about causal interactions are being encoded in parameters reflecting functional form and integration rules. Although these are certainly important steps in the direction of encoding mechanism knowledge, mechanism and force theories suggest that we have deeper intuitions about underlying mechanisms which may also need to be expressed in causal model theories. Moreover, singular causation, the main focus of force theories, has been neglected by causal model theories, although there are some attempts to model singular causation (Cheng, 1993; Sloman et al., 2009; Griffiths & Tenenbaum, 2009).

There are also findings which are critical for causal model theories. Empirical evidence for the insufficiency of simple causal models comes from a study by Rehder and

Burnett (2005), who have developed a reasoning task which allowed for testing people's intuitions about the Markov condition, which is arguably the most fundamental feature of Bayes nets. In their experiments subjects had to rate the conditional probability of an effect's presence given the state of its cause C. The crucial manipulation was whether other effects of C were present or absent (i.e., common cause model). According to the Markov condition, participants' ratings should be invariant across these conditions. The probability of each effect should be only dependent on its cause, not the states of the other effects, which are screened off by this cause. Contrary to this prediction, the ratings were clearly sensitive to the states of other effects of C. The more collateral effects were present, the higher were the rating of the conditional probability of the target effect given the presence of C. This Markov violation was extremely robust across many cover stories and domains.

One possible explanation of this finding is that people bring to bear additional assumptions about hidden causal events and mechanisms. Possibly the absence of the collateral effects of the target cause led subjects to infer that something was wrong with the underlying mechanism, hence their uncertainty about the target effect. Although such knowledge can be modeled within Bayesian causal models when they are augmented with hidden variables and mechanisms (Rehder & Burnett, 2005), the robustness of the finding indicates that the usually postulated minimal versions of causal models (e.g., Fig. 54.2) do not sufficiently capture our causal intuitions.

General Summary and Perspectives for Research

Our overview of research on causal reasoning has demonstrated how much progress has been made in this field in the past two decades. Not long ago theories of causal reasoning dominated which tried to get away with subsuming causal reasoning as a special case of more general theories of reasoning, thus ignoring the unique features of causality. In these theories causal reasoning has been viewed as a special case of

associative, logical, or probabilistic reasoning with minimal constraints coming from causal notions. Although domain-general reasoning certainly plays a role in causal domains, it has been shown that these theories fail to capture truly causal reasoning. People are sensitive to various aspects of causality including the directionality of the causal arrow, the structure of causal models, causal power, or forces and mechanisms, and this sensitivity needs to be an integral part of theories of causal reasoning. The present review has discussed a number of recent theories that capture the unique features of causality.

We only could touch upon a subset of studies on causal learning and reasoning. An increasing number of studies are interested in the relationship of causal reasoning with other cognitive tasks, including diagnostic reasoning (Fernbach, Darlow, & Sloman, 2010; Krynski & Tenenbaum, 2007; Meder, Mayrhofer, & Waldmann, 2009b), legal reasoning (Lagnado & Harvey, 2008), scientific explanations (Lombrozo, 2007, 2010), or analogical problem solving (Holyoak, Lee, & Lu, 2010; Lee & Holyoak, 2008). Moreover, the interaction between category learning and causal induction has been an important recent area of research (Kemp, Goodman, & Tenenbaum, 2010; Lien & Cheng, 2000; Marsh & Ahn, 2009; Waldmann & Hagmayer, 2006; Waldmann, Meder, von Sydow, & Hagmayer, 2009).

We have discussed research in the context of different theoretical paradigms, which have precursors in philosophy and Artificial Intelligence research. This leads to the obvious question how the different approaches can be unified. Obviously each theoretical paradigm has its pet empirical paradigm for which it works best, whereas other applications are neglected. Currently promising approaches are the attempts to integrate our knowledge about mechanisms with theories that pick up covariation information to induce causal power and causal structures (e.g., causal model theories). However, the

postulated mechanism knowledge is currently far more abstract than what force theories, for example, postulate.

The different levels of abstraction in the competing theories may turn out to be a blessing in disguise. It seems highly implausible that people represent causal relations uniformly at the same level of abstraction (e.g., as forces or as abstract causal arrows). Cartwright (2004) has expressed skepticism about the possibility of reducing causal relations to a few abstract concepts or to a uniform theory. Causal knowledge can be represented on an abstract level which is sufficiently captured by nodes and arrows in a Bayes net (e.g., "IQ influences motivation"), or can make very specific references to various mechanisms, which require more detailed representations (e.g., "the sun attracts the planets; "pistons compress air in the carburetor chamber"). Often, we may not care about mechanisms. If we intervene in a complex system, such as our economy, we are generally only interested in global outcomes rather than the myriads of arbitrary causal processes that govern complex systems. At other times, we spend huge efforts searching for the underlying mechanisms (e.g., air plane crashes).

One reason for shifting levels of abstraction is that causal explanations are sometimes more stable on the specific process level, sometimes more on the abstract level. For example, when we send off an e-mail we are fairly sure that the recipient will receive it, although it is unpredictable what path the message will take in the internet (see also Cartwright, 2001; Lombrozo, 2010). It seems plausible that various factors, including expertise and goals, influence how we choose to represent and use causal knowledge. This flexibility may in part justify why we have so many theories of causal reasoning.

References

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.

Ahn, W.-K., Kim, N. S., Lassaline, M. E., & Dennis, M. J. (2000). Causal status as a determinant of feature centrality. *Cognitive Psychology, 41,* 361-416.

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

Beckers, T., De Houwer, J., Pineño, O., & Miller, R. R. (2005). Outcome additivity and outcome maximality influence cue competition in human causal learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31*, 238-249.

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

Booth, S. L., & Buehner, M. J. (2007). Asymmetries in cue competition in forward and backward blocking designs: Further evidence for causal model theory. *Quarterly Journal of Experimental Psychology, 60*, 387-399.

Buehner, M. J. (2005). Contiguity and covariation in human causal inference. *Learning and Behavior, 33,* 230-238.

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 University Press, Cambridge, UK.

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

Buehner, M. J., & May, J. (2003). Rethinking temporal contiguity and the judgment of causality: Effects of prior knowledge, experience, and reinforcement procedure. *Quarterly Journal of Experimental Psychology, 56A,* 865-890.

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.

Carey, S. (2009). *The origin of concepts.* Oxford: Oxford University Press.

Cartwright, N. (1999). *The dappled world. A study of the boundaries of science.* Cambridge: Cambridge University 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.

Chapman, G. B., & Robbins, S. J. (1990). Cue interaction in human contingency judgment. *Memory & Cognition, 18*, 537-545.

Chapman, L. J., & Chapman, J. P. (1967). Genesis of popular but erroneous psychodiagnostic observations. *Journal of Abnormal Psychology, 72*, 193-204.

Chapman, L. J., & Chapman, J. P. (1969). Illusory correlation as an obstacle to the use of valid diagnostic signs. *Journal of Abnormal Psychology, 74*, 271-280.

Cheng, P. W. (1993). Separating causal laws from casual facts: Pressing the limits of statistical relevance. In D. L. Medin (Ed.), *The psychology of learning and motivation*, vol. 30 (pp. 215-264). New York: Academic Press.

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

Cheng, P. W., & Novick, L. R. (1991). Causes versus enabling conditions. *Cognition, 40*, 83-120.

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

Cheng, P. W., & Novick, L. R. (2005). Constraints and nonconstraints in causal learning: Reply to White (2005) and to Luhmann and Ahn (2005). *Psychological Review, 112*, 694-706.

Cheng, P. W., Novick, L. R., Liljeholm, M., & Ford, C. (2007). Explaining four psychological asymmetries in causal reasoning: Implications of causal assumptions for coherence. In M. O'Rourke (Ed.), *Topics in contemporary philosophy* (Vol. 4, pp. 1-32): *Explanation and causation*. Cambridge, MA: MIT Press.

Cummins, D. D. (1995). Naïve theories and causal deduction. *Memory & Cognition, 23*, 646-658.

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

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

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

Fenker, D. B., Waldmann, M. R., & Holyoak, K. J. (2005). Accessing causal relations in semantic memory. *Memory & Cognition*, *33*, 1036-1046.

Fernbach, P. M., & Sloman, S. A. (2009). Causal learning with local computations. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35*, 678-693.

Fernbach, P. M., Darlow A., & Sloman, S. A. (2010). Neglect of alternative causes in predictive but not diagnostic reasoning. *Psychological Science, 21*, 329-336.

Fugelsang, J., & Thompson, V. (2003). A dual-process model of belief and evidence interactions in causal reasoning. *Memory & Cognition, 31*, 800-815.

Goedert, K. M., Harsch, J., & Spellman, B. A. (2005). Discounting and conditionalization: Dissociable cognitive processes in human causal inference. *Psychological Science*, *16*, 590-595.

Goldvarg, E., & Johnson-Laird, P. N. (2001). Naïve causality: a mental model theory of causal meaning and reasoning. *Cognitive Science, 25*, 565-610.

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

Gopnik, A., Glymour, C., Sobel, D. M., Schulz, L. E., Kushnir, T., & Danks, D. (2004). A theory of causal learning in children: Causal maps and Bayes nets. *Psychological Review, 111,* 1-30.

Greville, W. J., & Buehner, M. J. (2007). The influence of temporal distributions on causal induction from tabular data. *Memory & Cognition, 35*, 444-453.

Greville, W. J., & Buehner, M. J. (2010). Temporal predictability facilitates causal learning. *Journal of Experimental Psychology: General, 139*, 756-771.

Griffiths, T. L., & Tenenbaum, J. B. (2005). Structure and strength in causal induction. *Cognitive Psychology, 51,* 354-384.

Griffiths, T. L., & Tenenbaum, J. B. (2007). Two proposals for causal grammars. In A. Gopnik & L. Schulz (Eds.), *Causal learning: Psychology, philosophy, and computation.* Oxford: Oxford University Press.

Griffiths, T. L., & Tenenbaum, J. B. (2009). Theory-based causal induction. *Psychological Review, 116,* 661-716.

Griffiths, T. L., Kemp, C., and Tenenbaum, J. B. (2008). Bayesian models of cognition. In R. Sun (Ed.), *The Cambridge handbook of computational cognitive modeling* (pp. 59-100). Cambridge University Press.

Hagmayer, Y., & Sloman, S. A. (2009). Decision makers conceive of their choices as interventions. *Journal of Experimental Psychology: General, 138*, 22-38.

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

Hagmayer, Y., & Waldmann, M. R. (2007). Inferences about unobserved causes in human contingency learning. *Quarterly Journal of Experimental Psychology, 60*, 330-355.

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

Hattori, M., & Oaksford, M. (2007). Adaptive non-interventional heuristics for covariation detection in causal induction: Model comparison and rational analysis. *Cognitive Science, 31,* 765-814.

Holyoak, K. J., Lee, H. S., & Lu, H. (2010). Analogical and category-based inference: A theoretical integration with Bayesian causal models. *Journal of Experimental Psychology: General, 139,* 702–727.

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

Jenkins, H. M., & Ward, W. C. (1965). Judgment of contingency between responses and outcomes. *Psychological Monographs, 79* (whole volume X).

Joyce, J. M. (1999). *The foundations of causal decision theory*. Cambridge, UK: Cambridge University Press.

Kelley, H. H. (1973). The process of causal attribution. *American Psychologist, 28,* 107-128.

Kemp, C., Goodman, N., & Tenenbaum, J. B. (2010). Learning to learn causal models. *Cognitive Science,* 34, 1185-1243.

Kim, N. S., & Ahn, W. (2002). Clinical psychologists' theory-based representations of mental disorders predict their diagnostic reasoning and memory. *Journal of Experimental Psychology: General, 131,* 451-476.

Koller, D., & Friedman, N. (2009). *Probabilistic graphical models: Principles and techniques.* Cambridge: MIT Press.

Koslowski, B. (1996). *Theory and evidence: The development of scientific reasoning.* Cambridge, MA: MIT Press.

Krynski, T. R. and Tenenbaum, J. B. (2007). The role of causality in judgment under uncertainty. *Journal of Experimental Psychology: General, 136,* 430-450.

Kushnir, T., Gopnik, A., Lucas, C., & Schulz, L. E. (2010). Inferring hidden causal structure. *Cognitive Science, 34,* 148-160.

Lagnado, D. A. & Harvey. N. (2008). The impact of discredited evidence. *Psychonomic Bulletin & Review.*

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

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

Lee, H. S., & Holyoak, K. J. (2008). The role of causal models in analogical inference. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34,* 1111-1122.

Leising, K. J., Wong, J., Waldmann, M. R., & Blaisdell, A. P. (2008). The special status of actions in causal reasoning in rats. *Journal of Experimental Psychology: General, 137,* 514-527.

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

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

Lombrozo, T. (2007). Simplicity and probability in causal explanation. *Cognitive Psychology, 55,* 232-257.

Lombrozo, T. (2010). Causal-explanatory pluralism: how intentions, functions, and mechanisms influence causal ascriptions. *Cognitive Psychology, 61,* 303-332.

López, F. J., & Shanks, D. R. (2008). Models of animal learning and their relations to human learning. In R. Sun (Ed.), *Cambridge handbook of computational psychology* (pp. 589-611). Cambridge, UK: Cambridge University Press.

López, F. J., Cobos, P. L., & Caño, A. (2005). Associative and causal reasoning accounts of causal induction: Symmetries and asymmetries in predictive and diagnostic inferences. *Memory & Cognition, 33*, 1388-1398.

Lu, H., Yuille, A. L., Liljeholm, M., Cheng, P. W., & Holyoak, K. J. (2008). Bayesian generic priors for causal learning. *Psychological Review, 115*, 955-982.

Lucas, C. G., & Griffiths, T. L. (2010). Learning the form of causal relationships using hierarchical Bayesian models. *Cognitive Science, 34*, 113-147.

Luhmann, C. C., & Ahn, W. (2007). BUCKLE: A model of unobserved cause learning. *Psychological Review. 114*, 657-677.

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

Mackay, D. J. C. (2003), Information theory, inference, and learning algorithms. Cambridge: Cambridge University Press.

Markovits, H., & Potvin, F. (2001). Suppression of valid inferences and knowledge structures: The curious effect of producing alternative antecedents on reasoning with causal conditionals. *Memory & Cognition, 29*, 736-744.

Marsh, J. K. & Ahn, W. (2009). Spontaneous assimilation of continuous values and temporal information in causal induction. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35*, 334-352.

McCloskey, M. (1983). Naive theories of motion. In D. Gentner & A. L. Stevens (Eds.), *Mental models* (pp. 299-324). Hillsdale: Erlbaum.

Meder, B., Hagmayer, Y., & Waldmann, M. R. (2008). Inferring interventional predictions from observational learning data. *Psychonomic Bulletin & Review, 15*, 75-80.

Meder, B., Hagmayer, Y., & Waldmann, M. R. (2009a). The role of learning data in causal reasoning about observations and interventions. *Memory & Cognition, 37*, 249-264.

Meder, B., Mayrhofer, R., & Waldmann, M. R. (2009b). A rational model of elementary diagnostic inference. In N. Taatgen & H. van Rijn (Eds.), *Proceedings of the Thirty-First Annual Conference of the Cognitive Science Society* (pp. 2176-2181). Austin, TX: Cognitive Science Society.

Medin, D. L., Coley, J. D., Storms, G., & Hayes, B. K. (2003). A relevance theory of induction. *Psychonomic Bulletin & Review, 10*, 517-532.

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

Neys, W., Shaeken, W., & d'Ydewalle, G. (2002). Causal conditional reasoning and semantic memory retrieval: a test of the semantic memory framework. *Memory & Cognition, 30*, 908-920.

Neys, W., Shaeken, W., & d'Ydewalle, G. (2003). Causal conditional reasoning and strength of association: the disabling condition case. *European Journal of Cognitive Psychology, 15*, 161-176.

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

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

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

Perales, J. C., & Shanks, D. R. (2007). Models of covariation-based causal judgment: A review and synthesis. *Psychonomic Bulletin and Review, 14*, 577-596.

Perales, J. C., Shanks, D. R., & Lagnado, D. (2010). Causal representation and behavior: The integration of mechanism and covariation. *Open Psychology Journal, 3,* 174-183.

Quinn, W. S., & Markovits, H. (1998). Conditional reasoning, causality, and the structure of semantic memory: strength of association as a predictive factor for content effects. *Cognition, 68*, B93-B101.

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

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

Rehder, B. (2009). Causal-based property generalization. *Cognitive Science, 33*, 301-343.

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

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

Rehder, B., & Kim, S. (2006). How causal knowledge affects classification: A generative theory of categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 32*, 659-683.

Reichenbach, H. (1956). *The direction of time*. Berkeley and Los Angeles: University of California Press.

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

Rozenblit, L., & Keil, F. C. (2002). The misunderstood limits of folk science: An illusion of explanatory depth. *Cognitive Science, 26,* 521-562.

Russell, Bertrand (1912/1992). On the notion of cause. In J. Slater (Ed.), *The collected papers of Bertrand Russell v6: Logical and philosophical papers 1909-1913* (pp. 193-210). London: Routledge Press.

Schlottmann, A., & Shanks, D. R. (1992). Evidence for a distinction between judged and perceived causality. *Quarterly Journal of Experimental Psychology: Human Experimental Psychology, 44(A),* 321-342.

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

Shanks, D. R., & Darby, R. J. (1998). Feature- and rule-based generalization in human associative learning. *Journal of Experimental Psychology: Animal Behavior Processes, 24,* 405-415.

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

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

Sloman, S. A., & Lagnado, D. A. (2005). Do we "do"? *Cognitive Science, 29,* 5-39.

Sloman, S. A., Barbey, A. K., & Hotaling, J. (2009). A causal model theory of the meaning of “cause,” “enable,” and “prevent.” *Cognitive Science, 33,* 21-50.

Sobel, D. M., Tenenbaum, J. B., & Gopnik, A. (2004). Children's causal inferences from indirect evidence: Backwards blocking and Bayesian reasoning in preschoolers. *Cognitive Science, 28,* 303-333.

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.

Talmy, L. (1988). Force dynamics in language and cognition. *Cognitive Science, 12*, 49-100.

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. (2007). Combining versus analyzing multiple causes: How domain assumptions and task context affect integration rules. *Cognitive Science, 31,* 233-256.

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., & Hagmayer, Y. (2006). Categories and causality: The neglected direction. *Cognitive Psychology, 53,* 27-58.

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., & Walker, J. M. (2005). Competence and performance in causal learning. *Learning & Behavior, 33*, 211-229.

Waldmann, M. R., Cheng, P. W., Hagmayer, Y., & Blaisdell, A. P. (2008). Causal learning in rats and humans: a minimal rational model. In N. Chater, & M. Oaksford (Eds.), *The probabilistic mind. Prospects for Bayesian Cognitive Science* (pp. 453-484). Oxford: University Press.

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.

Waldmann, M. R., Meder, B., von Sydow, M., & Hagmayer, Y. (2009). The tight coupling between category and causal learning. *Cognitive Processing, published online.*

White, P. A. (2003). Making causal judgments from contingency information: The pCI rule. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29*, 710–727.

White, P. A. (2005). Postscript: differences between the causal powers theory and the power PC theory. *Psychological Review, 112***,** 683-684.

White, P. A. (2006). The causal asymmetry. *Psychological Review, 113***,** 132-147.

White, P. A. (2009). Perception of forces exerted by objects in collision events. *Psychological Review, 116*, 580-601.

Wolff, P. (2007). Representing causation. *Journal of Experimental Psychology: General, 136*, 82-111.

Wolff, P., & Song, G. (2003). Models of causation and the semantics of causal verbs. *Cognitive Psychology, 47*, 276-332.

Wolff, P., Barbey, A. K., & Hausknecht, M. (2010). For want of a nail: How absences cause events. *Journal of Experimental Psychology: General*, *139*, 191-221.

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.