KNOWLEDGE-BASED CAUSAL INDUCTION

Michael R. Waldmann

I. Introduction

Our ability to acquire causal knowledge is central for our survival. Causal knowledge allows us to predict future events and to plan actions to achieve goals. The importance of causal knowledge is the reason why this topic has attracted many philosophers and psychologists in the past. Philosophical analyses tend to focus on the ontological characteristics of causality, whereas psychological theories are primarily interested in the processes of acquiring and representing causal knowledge. Despite this apparent division of labor, the two approaches are strongly connected. For example, David Hume, one of the forefathers of modern views on causality, claimed that everything we possibly know about the causal texture in the world is based on associations between perceived events (Hume, 1739/1978; 1748/1977). This view has proven extremely influential. It still dominates modern psychological thinking on causality. However, many of Hume's insights, which have been preserved in modern *philosophical* analyses, have actually been lost in current *psychological* theories that tried to reconcile Hume's view with modern psychological learning theories.

II. The Associative View

A. Hume's Heritage

David Hume has influenced modern thinking about causality more than other philosophers (see Mackie, 1974). Hume may be viewed as the fore-

THE PSYCHOLOGY OF LEARNING
AND MOTIVATION, VOL. 34

father of modern psychological theories that try to reduce causal knowledge to associative links (see Shanks, 1993; Shanks & Dickinson, 1987; Wasserman, 1990; Young, 1995). Hume postulated three types of associations: (1) resemblance; (2) spatiotemporal contiguity; and (3) cause–effect relations. Thus, he clearly differentiated between associations that are based on spatiotemporally contiguous single events, and those based on cause and effect. Causal associations are accompanied by the impression of a "necessary connexion." One of Hume's main interests was the question of what this impression is based on. The traditional answer that forms the background of Hume's theory postulated that causal impressions are based on the observation of *causal powers* that are transmitted from causes to effects. By contrast, Hume, being an Empiricist, assumed that all our reasoning is based on the observation of singular separated events. This ontological framework made it impossible for him to find anything like causal processes or powers. Instead he concluded that the impression of a necessary connection between causes and effects is actually a *cognitive illusion* based on an associative relation (i.e., "habit," "custom") that is caused by *repeated* observations of paired events. According to Hume (1739/1978), causal impressions are formed when the following three constraints are met:

(1) The cause and effect must be contiguous in space and time.
(2) The cause must be prior to the effect.
(3) There must be a constant union betwixt the cause and effect. "Tis chiefly this quality, that constitutes the relation." (p. 173)

Hume, particularly in his later work (Hume, 1748/1977), did not deny that causal powers may exist in the world. However, he insisted that we are unable to observe causal powers directly. Our causal impressions are based on the strength of associative links. Like his modern successors (see Wasserman, 1993), Hume thought that the importance of causal knowledge for our survival is the reason why causal impressions are based on low-level mechanical processes rather than higher order reasoning:

> It is more comfortable to the ordinary wisdom of nature to secure so necessary an act of mind, by some instinct or mechanical tendency, which may be infallible in its operations, may discover itself at the first appearance of life and thought, and may be independent of all the laboured deductions of the understanding. (Hume, 1748/1977, p. 37)

B. From Stimulus–Response Learning to Causal Induction

Even though Hume's philosophy may be viewed as a predecessor of modern learning theories, the adoption of Hume's theory of causal induction is a relatively late achievement, and did not occur without costs. The traditional, behavioristically oriented learning theories viewed learning as the acquisition of associative links between stimuli and response (e.g., Pavlov, 1927), or behavior and outcomes (e.g., Thorndike, 1911). In the past 20 years, a new cognitive view of associative learning emerged that bears much more resemblance to Hume's view than to traditional reflex-oriented theories. This approach can be traced back to Tolman and Brunswik's (1935) work in which they argued that the primary goal of learning is the discovery of the causal texture of the world. Mackintosh (1983) summarizes the modern view:

> The suggestion, then, is that as a result of conditioning animals acquire knowledge about their environment which maps the relationship between events occurring in that environment. The function of conditioning, it has been suggested, is precisely to enable animals to discover the causal structure of their world. . . . (p. 11)

According to this view, Pavlov's dogs learned to predict food on the basis of the tone cue, rather than simply strengthening a reflex between the cue and the salivating response. However, in the process of translating Hume's causes and effects into the behaviorist language of stimuli and responses, some of Hume's insights were lost. Most notably, Hume's conceptual distinction between causes and effects that is reflected in his temporal priority assumption was dropped when the cue–outcome terminology of early reflex psychology was preserved. Unlike Hume, modern psychological theories of associative learning typically describe learning as the acquisition of associative links between *cues* and *outcomes* rather than causes and effects.[1] Most saliently, associative theories still use the terms "conditioned stimuli" (CS) and "unconditioned stimuli" (US) when describing human and animal causal learning.

The reduction of causes and effects to cues and outcomes is motivated by the behaviorist background assumptions of psychological associationism. The organism is conceived of as responding to the actual stimuli regardless of what type of events these stimuli actually represent. Cues play a double causal role. On one hand, they represent events in the outside world. These events may be causes or effects. On the other hand cues *cause* responses, sometimes via the representation of intermediate steps. The associative

[1] In order to clarify the difference between causal-model theory and associative theories, a generic paradigmatic case of associative learning theory is discussed here. This chapter focuses on associative theories of human causal induction, not on associative learning in general. In this area, the Rescorla–Wagner theory currently dominates (Rescorla & Wagner, 1972), and will therefore primarily be discussed. However, some alternative associative theories that have been proposed in the context of human causal induction will also be discussed briefly.

links between cues and outcomes reflect the strength of this *internal* causal effectiveness of cues. Typically the strength of associative weights is conceived of as representing the organism's assessment of the strength of causal relations between causes and effects. However, whether or not the internal causal relation between cues and outcomes reflects causal relations between actual causes and effects rather than other types of event relations is simply a matter of coincidence.

This reductionism to a nonrepresentational theory of learning (see also Gallistel, 1990) about causal relations is already apparent in Tolman and Brunswik's (1935) theory. Object perception, for example, is described as based on the causal relation between the distal stimulus, the object, and the proximal stimulus, the cue. The cue is the effect of the object. However, this causal relationship is lost when the authors switch from the description of the outside world to their psychological theory of object perception. This process is described as involving a process of cue integration irrespective of the causal role of the events corresponding to the cues.

Sometimes it has been implicitly assumed that cues (CS) typically code causes and outcomes (US) code effects (e.g., Van Hamme, Kao, & Wasserman, 1993; Esmoris-Arranz, Miller, & Matute, 1995), so that "CS" is just a shorthand for "cause," and "US" for "effect." However, this correspondence holds only for learning situations in which the organism is presented with causes as the learning input, when it generates a prediction parallel to the unfolding of the causal processes in the world, and then compares its prediction with the observed effects. Not all learning is stimulus bound in this sense. Learning situations may be constructed in which the information processing system responds to effect cues, and tries to figure out the causes of these effects (e.g., Waldmann & Holyoak, 1992). In this situation the internal causes of the response, the cues, correspond to effects in the outside world. Thus, the causal relations expressed by the associative links that trigger the response do not reflect the causal relations of the corresponding events in the world.

In summary, associative learning theories view causal induction as a data-driven process in which causes and effects are represented as cues and outcomes. Learning involves the acquisition of associative links between cues and outcomes. The primary role of these links is the elicitation of outcome representations. These links may reflect causal relations between causes and effects in situations in which cues actually represent causes. However, this correspondence is not a consequence of associative weights actually *representing* causal relations; it is rather a fortuitous coincidence of a restricted set of learning situations.

III. Causal-Model Theory

The majority of theories of causal induction focus on bottom-up processes of knowledge acquisition (e.g., Anderson, 1990; Cheng, 1993; Shanks & Dickinson, 1987). Typically the potential impact of domain-specific knowledge on the learning process is acknowledged but it is argued that learning can be studied separately from knowledge influencing the learning process. Associative theories, for example, may accommodate knowledge influences by assuming that in some learning situations the learning process starts with associative weights that have been transferred from previous learning occasions (see Alloy & Tabachnik, 1984; Choi, McDaniel, & Busemeyer, 1993). Similarly, the probabilistic contrast theory (Cheng, 1993; see also Cheng, Park, Yarlas, & Holyoak, this volume, Ch. 8) focuses on data-driven processes that generate causal knowledge as the output of the processing of statistical contingency information. The acquired knowledge may then be the basis of further learning. Thus, both research paradigms assume that causal knowledge is primarily acquired by means of bottom-up processes. This knowledge may then later affect learning, but bottom-up acquisition of causal knowledge and top-down influences are viewed as two processes that can be studied separately and independent of each other.

By contrast, causal-model theory (Waldmann & Holyoak, 1992; Waldmann, Holyoak, & Fratianne, 1995) assumes that the acquisition of causal knowledge is characterized by an interaction of data-driven and knowledge-driven processes (see also Wisniewski & Medin, 1994, for a similar view). This view is compatible with many findings that demonstrate the impact of *domain-specific* knowledge on learning (see Murphy & Medin, 1985). However, causal-model theory pursues the more ambitious goal of demonstrating that knowledge also influences learning in situations in which no prior domain-specific knowledge is available. It is assumed that in these situations more *abstract* kinds of knowledge are activated. Causal-model theory generally claims that causal induction is guided by knowledge. Causal models provide the basis for the interpretation of the learning input. The "tight coupling" (Wisniewski & Medin, 1994) between the learning input and top-down interpretations is the reason why knowledge and learning cannot be studied separately. The assumption of a necessary interaction between experience and abstract knowledge in the process of knowledge acquisition can be traced back to Kant's (1781/1950) philosophy. Kant postulated in his criticism of Empiricist philosophies that knowledge is possible only when the sensory input is interpreted by a priori categories of knowledge. Even though causal-model theory does not subscribe to Kant's particular view on causality (see Mackie's, 1974, critical review), its general tenet that the learning input interacts with interpretative processes is in the spirit of Kant's epistemology.

Causal models provide the basis for the flexible interpretation of the learning input. Unlike in associative theories, the learning cues can be assigned flexibly to represent causes or effects in the causal representation of the learning situation. Causal-model theory postulates that causal induction attempts to arrive at adequate *representations* of the world regardless of the order in which information about the constituents of these representations is acquired.

A. Causal Directionality

One of the most important examples of abstract causal knowledge that may affect the processing of the learning input is knowledge about causal directionality. We know that the causal arrow is directed from causes to their effects and not the other way around. This fundamental property of causal relations is of the utmost pragmatic importance as it provides the basis for our ability to reach goals. Effects can be achieved by manipulating causes but causes cannot be accomplished by manipulating their effects. Thus, it is extremely important to be able to distinguish between causes and effects.

Accounting for causal asymmetry presents a problem for many philosophical theories of causality. Bromberger (1966) criticized Hempel's (1965) seminal theory of deductive-nomological explanation using the example of a flagpole: We can explain the length of a shadow cast by a flagpole by premises that include a statement about the length of the flagpole, the elevation of the sun, and the laws of the propagation of light. But, equally, we can derive the height of the flagpole from the length of the shadow, the elevation of the sun, and the laws of the propagation of light. Both are perfect examples of deductive-nomological explanations. Therefore, this scheme does not account for the fundamental property of causal asymmetry. Similarly, theories characterizing causes as necessary and/or sufficient conditions of their effects fail in this regard, since it is equally true that effects are necessary and/or sufficient conditions of their causes (Mackie, 1974; von Wright, 1971).

Probabilistic theories of causality represent a more recent approach (Eells, 1991; Salmon, 1971; Suppes, 1970). Roughly, it has been proposed that *causes alter the probabilities of their effects.* This idea has been adopted by psychologists who propose that causal induction involves the acquisition of knowledge about *contingencies* between causes and effects (Cheng & Novick, 1992; Cheng et al., this volume, Ch. 8; Jenkins & Ward, 1965; Pearl, this volume, Ch. 10; Wasserman, Chatlosh, & Neunaber, 1983). Formally, an (unconditional) contingency (Δp) can be defined as the difference between the conditional probability of a target effect E given the presence of a potential causal factor C and its probability given the absence of the factor ($\sim$C), that is,

$$\Delta p = p(\mathrm{E}|\mathrm{C}) - p(\mathrm{E}|\sim\mathrm{C}). \quad (1)$$

This formula allows for the representation of positive, excitatory causes ($\Delta p > 0$) and negative, inhibitory causes ($\Delta p < 0$). Contingencies per se also do not account for causal asymmetry. The problem arises from the fact that statistical correlations are symmetric. When a cause raises the probability of its effect, the reverse is typically also true, namely that the effect raises the probability of its cause.

Finally, associative accounts also fail to reflect the priority of causes. In most theories, associative weights are directed from cues to outcomes regardless of whether the cues represent causes or effects (see Waldmann & Holyoak, 1992).

In order to account for causal directionality, philosophical theories have typically followed Hume's lead and have included additional assumptions in their definitions of causality. Like Hume, many theorists added a criterion of temporal precedence as a basic characteristic of causal relations: causes temporally precede their effects (see Eells, 1991; Suppes, 1970). Psychologists who postulate that causal induction involves learning about statistical contingencies have also embraced this additional background assumption (Cheng & Novick, 1992; Einhorn & Hogarth, 1986).

Another frequently discussed criterion of causal directionality emphasizes the fact that the active manipulation of causes produces their effects but not the other way around (see Mackie, 1974; von Wright, 1971). Our ability to actively *intervene* in the processes taking place in the world allows us to impose a causal structure on the pattern of observed event covariations. The importance of our actions for our understanding of causality has also been elaborated by Piaget (1930).

A statistical method to distinguish between causes and effects has been proposed by the philosopher Reichenbach (1956, see also Pearl, 1988, this volume, Ch. 10; Salmon, 1984). In situations with multiple causes and multiple effects a typical statistical pattern emerges. Multiple correlated effects are rendered conditionally independent once their common cause is held fixed, but multiple causes cannot be rendered conditionally independent by holding fixed their common effect. A famous example involves a group of actors who suffer from a stomach disease after having dined together. Even though there is a small chance that this is a coincidence, the more plausible hypothesis is that food poisoning is the common cause of their illnesses. Conditional on the common cause of food poisoning, the individual illnesses are independent. As Reichenbach points out in his

principle of the common cause, coincidences, may be explained by a common cause but not by a common effect. According to Reichenbach, this typical statistical pattern is a characteristic feature of the physical world. Philosophical theories that model causality as the transmission of energy (Fair, 1979), conserved quantities (Dowe, 1992), or information (Salmon, 1984) derive this feature from the fundamental physical fact that the paths of multiple causes converging on a common effect meet, whereas multiple effects emerging from a common cause are reached on independent paths.

Finally, it has been proposed that causal directionality is based on the coherence of a postulated new causal relation with our general world knowledge (Kitcher, 1989). Postulating the flagpole as the cause of the shadow rather than the reverse certainly fits better with our prior knowledge about characteristics of physical objects.

Background assumptions about differences between causes and effects have often been implicitly invoked in psychological experiments even though they have rarely been explicitly acknowledged. The experiments of Wasserman and his colleagues may suffice as one example (see Wasserman, 1990). In a typical set of experiments, Wasserman et al. (1983) presented the participants in their experiments with the task of periodically pressing a key and subsequently observing the state of a light. Wassserman et al. found that the ratings of the causal effectiveness of the key pressing corresponded surprisingly well with the objective contingencies. This is a task in which no specific knowledge about the event relations was available. However, abstract knowledge may have helped to assign causal roles to the events so that the proper cause–effect contingencies could be computed. First, the causes (key pressing) occurred *temporally prior* to the effects (temporal priority criterion). The participants were requested to actively manipulate these causes by pressing keys (intervention criterion). The rating instructions suggested the causal roles of the events (instruction-based assignment), and, finally, interpreting key presses as causes and lights as effects is certainly more consistent with our world knowledge than other assumptions (coherence criterion). This is an example of how several of the criteria of causal directionality may converge.

Causal-model theory is based on the assumption that causal induction cannot solely be based on the processing of statistical information. Additional top-down assumptions, for example, about causal directionality, have to guide the processing of the learning input. In the next sections, a number of empirical studies are reported that show how abstract knowledge interacts with the processing of the learning input. The next section presents experiments that investigated the ease of acquiring different category structures; it is shown that prior assumptions about patterns of causal directionality influence the learning of otherwise identical learning inputs. Section IIIC focuses on one of the most important assumptions of current associative learning theories. The majority of these theories postulate cue competition. Experiments are presented that show that cue competition in fact interacts with the causal role of the cues. These results are a further demonstration of the impact of assumptions about causal directionality. In Section IIID a different example of the interaction between knowledge and learning is presented. It is shown that the way statistical contingencies between a putative cause and an effect are computed is influenced by background assumptions about the causal relevance of additional, potential cofactors This background knowledge has to be in place at the outset of the induction process in order to guide the acquisition of new knowledge. Section IIIE discusses the place of causal–model theory within the debate between theorists that view causal induction as based on the processing of statistical covariations and theorists who focus instead on causal mechanisms. A reconciliation between these two apparently different stances is offered. Section IIIF, finally, shows that not only prior assumptions about the causal role of the learning cues but also the order in which the learning input is presented may affect the causal representation of a learning situation.

B. Causal Models and the Learning of Category Structures

In order to test causal-model theory against associationist theories of categorization, Waldmann et al. (1995) designed a learning task in which participants received identical cues and had to learn identical outcomes, while the causal roles of the cues were varied. Standard associationist theories of categorization that model learning as the acquisition of associative weights between cues and outcomes would treat these tasks identically regardless of the causal status of the cues and the outcomes (e.g., Gluck & Bower, 1988a; Shanks, 1991; Shanks & Lopez, in press). By contrast, causal-model theory claims that participants should be sensitive to the causal roles of the cues and the outcomes, and to the different structural implications of the causal models that are used to interpret the learning input (see also Eddy, 1982; Tversky & Kahneman, 1980).

Figure 1 depicts the two causal models that were used in the experiments of Waldmann et al. (1995). Both models consist of four elements but the causal directions connecting these elements entail distinct covariational patterns. Figure 1A shows a *common-cause* structure in which a common cause simultaneously produces three effects. Figure 1B shows a *common-effect* structure in which three causes independently produce a single effect. A key difference between these two structures is that common-cause structures imply a spurious correlation among their effects. Even though the

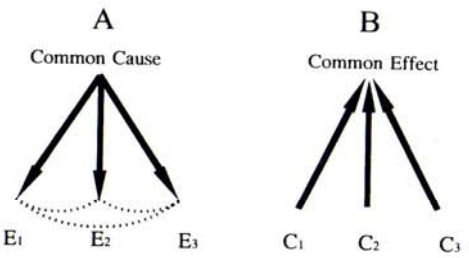

Fig. 1. Common-cause structure (A) with multiple independent effects (E_1,E_2,E_3) versus common-effect structure (B) with independent causes (C_1,C_2,C_3). Only the common-cause structure formally implies a spurious correlation (dotted curves) among effects. From Waldmann et al. (1995). Copyright © 1995 by the American Psychological Association. Reprinted with permission.

effects do not affect each other, they tend to covary as the status of the common cause varies. In contrast, a common-effect structure does *not* imply a correlation among its causes. It is possible that several causes may interact, but in such cases the underlying causal model has to be augmented to account for these interactions. The need to modify the causal model by adding explicit configural features would be expected to increase the difficulty of learning (Dawes, 1988). In general, causal-model theory predicts that learning difficulty should be dependent on the fit between the structural implications of the causal models activated during learning and the structure of the learning input.

Figure 2 displays an example of the learning materials of Experiment 4 of Waldmann et al. (1995). The cards showed stones in the middle surrounded by three colored iron compounds. The task was to judge whether the stone in the middle of the dish was a magnet or not.

All participants saw the same pictures with the stones and the iron compounds. However, we used two different instructions, which manipulated the direction of the causal arrow connecting stones and compounds. In the *common-cause* context, participants were told that scientists had discovered that some of these stones are either strong or weak magnets. In order to find out more about these stones, the scientists put the stones in dishes along with iron compounds. They found out that stones that are magnetic change the orientation of some of the iron compounds placed in the dish. Strong magnets turn the magnetized compounds so that their ends point to the stone, weak magnets turn the magnetized compounds so that their sides face the stone. If the stones are not magnetic, the iron compounds just stay in a random orientation. The participants' task was to learn to judge whether a stone was a magnet or not, basing their decisions on the

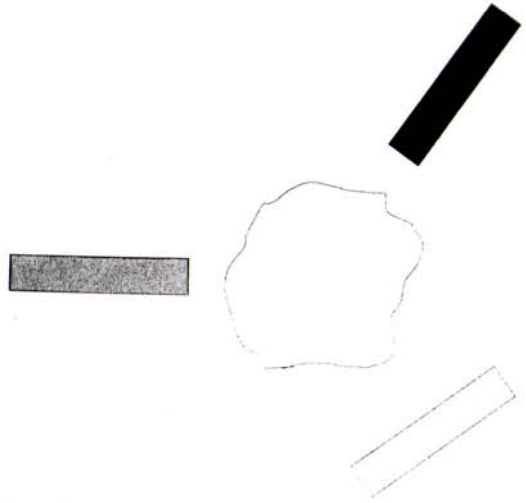

Fig. 2. Example of the learning material from Waldmann et al.'s (1995) Experiment 4. A potential magnet surrounded by iron compounds (compounds were blue, red, and green in the original set of learning items). From Waldmann et al. (1995). Copyright © 1995 by the American Psychological Association. Reprinted with permission.

orientation of the surrounding compounds. No prior information was given about which of the different compounds were actually affected by the magnets. Participants were presented with individual cases one after the other, they had to decide whether they believed the stone displayed on the index card represented a magnet or not, and subsequently were informed whether they were correct or not. Thus, no feedback about whether the magnet was strong or weak was provided.

In the common-effect conditions, the same material was used but in the initial instructions the direction of the causal connections between stones and compounds was reversed. In these conditions participants were told that scientists had discovered that some of the iron compounds emit strong or weak magnetic waves that may magnetize the stones and turn them into strong or weak magnets. The intensity of the magnetic waves was based on the orientation of these compounds: compounds pointing to the stone emit strong magnetic waves, whereas compounds facing the stones emit weak magnetic waves. Again, the participants' task was to learn to judge whether a stone was a magnet or not by using information about the orientation of the compounds surrounding the stone. Except for the differ-

ent initial instructions the learning procedure was identical across the two causal conditions.

In both causal situations the same entities are causally linked, only the direction of the causal arrow differs. In the common-cause instruction participants were confronted with a varying common cause, a strong or a weak magnet. This variation suggests that the affected compounds point to the stone when the cause is strong, and that their sides face the stone when it is weak. Thus, the common-cause model with a cause varying between a strong and a weak state should sensitize participants to a within-category correlation between the orientations of the affected compounds. These compounds should all be expected to either point to the stone (indicating a strong magnet) or face the stone (indicating a weak magnet). By contrast, the common-effect model with a varying effect does not structurally imply a within-category correlation between the causes. Here, it is more natural to assume three independent causes converging on a joint effect. We expected that the common-effect instruction should sensitize participants to category structures that exhibit independent cue-to-category correlations.

To test these predictions, we presented participants with either a category structure that embodies cue-to-category correlations, or a structure that contains a within-category correlation. Causal-model theory predicts that participants in the common-cause conditions should be biased to expect a within-category correlation, whereas participants in the common-effect conditions should find cue-to-category correlations more natural. Learning the within-category correlation after having received the common-effect instruction amounts to learning about a disordinal interaction among causes, which should be particularly hard to grasp. By contrast, the structure with cue-to-category correlations embodies a situation with three linear main effects within this causal condition.

More specifically, half of the participants received a linearly separable arrangement, which exhibits cue-to-category correlations. In this category structure, compounds pointing to the stones were more typical for the positive set ("yes"), and compounds parallel to the stones were more typical for the negative set ("no"). The other condition represented a non–linearly separable category structure in which the position of the individual compounds was not correlated with the categories. The only way to distinguish the two sets was by noticing the within-category correlation between two of the compounds. In the positive set, these two dimensions were perfectly positively correlated (i.e., both compounds either pointed to the stone or were positioned parallel to the stone); in the negative set they were negatively correlated (i.e., one of the two compounds pointed to the stone, the other compound was parallel to the stone). The non–linearly separable

structure corresponded to an Exclusive-Or (XOR) structure with an additional irrelevant feature.

Figure 3 displays the results of Experiment 4 (Waldmann et al., 1995). The mean number of errors until participants reached the learning criterion was used as an indicator of learning difficulty. Within the common-cause condition the non–linearly separable structure with the within-category correlation was easier to learn than the linearly separable structure with the cue-to-category correlations, whereas the opposite was true within the common-effect conditions. The interaction between causal condition and category structure proved highly reliable. The results support the view that participants are sensitive to the underlying causal structure of the task domain. Since across the two causal conditions participants saw identical cues and had to learn to associate them with identical outcomes, these results cannot be explained by standard associationist theories that would generally assign the learning cues (compounds) to the input level and the outcomes (magnets) to the output level of an associationist network.

1. Configural Cues and the Learning of Causal Categories

A further problem for associationist theories such as the Rescorla–Wagner theory is the fact that the non–linearly separable structure proved learnable. It is a well-known fact that this theory is restricted to linearly separable

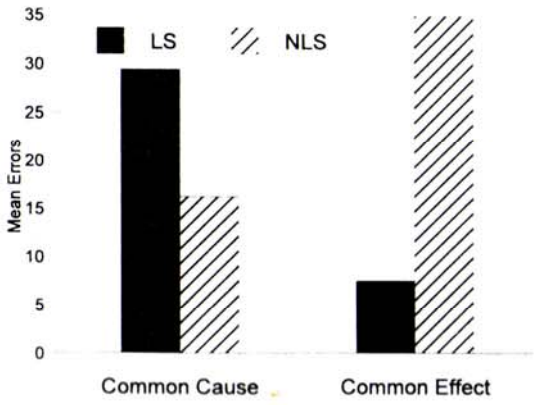

Fig. 3. Mean errors obtained in the linearly separable (LS) and the non–linearly separable (NLS) common-cause and common-effect conditions (Experiment 4 from Waldmann et al., 1995). The NLS conditions embody a within-category correlation, the LS conditions embody cue-to-category correlations.

tasks (Minsky & Papert, 1969). As a consequence, more complex theories have been suggested in which configural cues (Gluck & Bower, 1988b; Rescorla, 1973) or hidden layers (Kruschke, 1992; Rumelhart, Hinton, & Williams, 1986) are added to code interactions. However, even though these theories predict that nonlinear tasks are learnable they still do not account for the fact that the participants proved sensitive to the causal status of the cues. All these theories have in common that they try to associate cues with outcomes regardless of the causal structure connecting these events.

The configural-cue model in which cues coding conjunctions of elements are added to the input layer of an associationist network along with the elemental cues has an additional problem: the number of cues grows exponentially with the number of elemental input cues. Gluck and Bower (1988b) therefore suggested restricting configural cues to pairwise conjunctions. An obvious drawback of this restriction is that such a network is unable to handle problems for which the correct decision requires learning an interaction among three (or more) cues. Note that networks with hidden layers will also not necessarily learn all higher order interactions. If the number of hidden units is too small, the network might be able to learn a two-way interaction but not some higher order interaction (see Kruschke, 1992). One problem with many such learning networks, therefore, is that the complexity of the learning problem has to be anticipated in advance in order to pick the appropriate size of the network.

Waldmann and Holyoak (1990) expected that within a common-cause context with a varying cause a *three-way correlation* should be learned fairly easily because it falls out of a linear model with a common cause independently affecting three effects. In contrast, in a common-effect context with a varying effect a three-way interaction of three causes should be particularly difficult to grasp.

Waldmann and Holyoak (1990, Experiment 3) conducted an experiment in which participants received four cues that were characterized either as causes of a common effect or as effects of a common cause. Three of these cues were perfectly correlated within the positive set to which participants had to learn to respond with "yes." The experiment yielded two major results. First, the three-way interaction was clearly learnable, which refutes Gluck and Bower's (1988b) restrictive assumption on configural cues. Second, despite the fact that learning cues, response, and learning feedback were equated across the two causal conditions, a clear learning advantage for the common-cause condition was obtained (errors: $M = 43.1$ vs $M = 76.0$). Again, participants proved sensitive to the different structural implications derived from differential patterns of causal directionality.

C. Causal Models and Asymmetries of Cue Competition

Since Kamin (1969) discovered the phenomenon of *blocking* in animal learning, *cue competition* has been a basic phenomenon that all associative learning theories are trying to capture. In the classic blocking paradigm, animals are first (Phase 1) trained to associate an initial conditioned stimulus CS_1 with an unconditioned stimulus US. In Phase 2 of the learning procedure, a second cue CS_2 is redundantly paired with the initial cue CS_1. Kamin's crucial finding was that, in spite of being perfectly correlated with the outcome, the later redundant cue CS_2 did not seem to acquire any associative strength as compared to a control group, which did not receive any Phase 1 training.

Rescorla and Wagner's theory (1972) views blocking as the result of a failure to acquire associative strength. According to this rule learning is error driven. In blocking experiments, animals learn to predict the outcome using the initially acquired predictive cue CS_1. Since this cue still allows perfect predictions in Phase 2, no further learning occurs. In particular, the associative weight of CS_2 stays at its initial value of zero.

Waldmann and Holyoak (1992) modified the blocking paradigm in order to test causal-model theory against the Rescorla–Wagner and similar theories. As pointed out by Reichenbach (1956), one crucial characteristic of causal relations in the physical world is the fact that multiple independent causes of a common effect potentially interact, whereas multiple independent effects of a common cause are conditionally independent. Waldmann and Holyoak asked whether our learning is sensitive to this fundamental physical feature.

In a set of experiments, Waldmann and Holyoak (1992) employed a two-phase blocking design. In Phase 1 a predictive cue (P cue) was established as the sole deterministic predictor of an outcome (along with other nonpredictive cues). In Phase 2 this P cue was paired with a second, redundant predictor (R cue) as predictive of the outcome. The P and the R cues either always occurred together or were both absent. Two conditions were compared in which the causal status of the cues was manipulated by means of different initial instructions. Otherwise the two conditions presented exactly the same learning experiences. Thus, the interaction of blocking with the manipulation of the causal assumptions about the learning cues could be tested by comparing the results of these two groups. In the *predictive learning conditions* the cues were characterized as potential causes of a common effect. In Experiments 1 and 2 (Waldmann & Holyoak, 1992), for example, the cues were descriptions of the appearance of fictitious persons (e.g., "pale skin, stiff posture, normal perspiration"), and in the predictive condition these cues were described as potential causes of an

emotional response of observers of these persons. The crucial dependent measure in this condition was ratings of whether each of the cues represented an independent *cause* of the effect.

The Rescorla–Wagner rule predicts complete blocking of the R cue because it is redundantly paired with the P cue that was already established as perfectly predictive in Phase 1. A number of previous studies have demonstrated blocking with this kind of learning task (e.g., G. B. Chapman, 1991; G. B. Chapman & Robbins, 1990; Shanks, 1985).

Causal-model theory makes a similar prediction. Following recent developments of statistical relevance theory, Waldmann and Holyoak (1992) proposed that in situations with multiple causes converging on a common effect, *conditional contingencies* should be computed (see Cartwright, 1983; Cheng, 1993; Cheng & Novick, 1992; Eells, 1991; Melz, Chenz, Holyoak, & Waldmann, 1993; Salmon, 1980; Spellman, this volume, Ch. 5). Conditional contingencies (Δp_{K_i}) assess the contingencies between two events C and E conditional upon alternative causal factors K_i being kept constant, that is, as

$$\Delta p_{K_i} = p\ (E|C.K_1.K_2. \ldots K_n) - p(E|{\sim}C.K_1.K_2. \ldots K_n). \qquad (2)$$

An isolated period denotes an "and," and each K_i a choice between the presence or the absence of the factor. The computation of conditional contingencies is necessary to distinguish between true causal and spurious correlations. For example, suppose we want to test the hypothesis that smoking (C) causes lung cancer (E). Furthermore, we assume that smoking is correlated with alcohol consumption, which may also be a cause of lung cancer. In order to test the hypothesis, we should assess the conditional contingencies between smoking and lung cancer in the subpopulation of alcoholics (K_1) and people who do not drink alcohol ($\sim K_1$). If we then discover that smoking equally leads to lung cancer in both subpopulations, we may conclude that smoking is an independent cause of this disease.

A typical feature of the blocking design is the fact that conditional contingencies between the R cue and the effect cannot be computed in the absence of the P cue that has been established as an individual cause in Phase 1. The R cue is never presented alone without the P cue. Thus, causal-model theory predicts that the participants of the learning experiment should be uncertain as to whether the R cue represents a genuine cause or not. However, in contrast to the predictions of the Rescorla–Wagner theory, blocking is expected to be partial: Rather than concluding that the R cue is not a cause, participants should be uncertain, since they are simply not given crucial information, which is necessary to arrive at a definite assessment of the causal status of the R cue. The results of the experiments showed indeed that blocking was partial, as the ratings for the

R cue were substantially lower than those for the P cue, but higher than those for other cues that were uncorrelated with the effect (see also G. B. Chapman & Robbins, 1990).

In a second condition, the *diagnostic learning condition*, the very same cues of the predictive conditions were redefined as potential effects of a common cause. The participants were told that the persons' features represent potential effects of a new disease caused by a virus. Thus, in this condition the participants were confronted with a common-cause situation.

Causal-model theory claims that the participants honor the cause–effect direction regardless of the order of presentation of the components of the common-cause model. Since there is only one cause in common-cause situations, the conditional contingency rule (Eq. 2) reduces to unconditional contingencies (Eq. 1) between the single cause and the effects. Because both the P cue and the R cue are deterministic effects of the common cause (the virus), no blocking was predicted in the diagnostic condition. In the diagnostic condition of Experiment 1 (Waldmann & Holyoak, 1992) the participants were asked to rate the degree to which they thought each of the cues represented an independent *effect* of the cause. As predicted, no cue competition was found in this condition.

It is interesting to note that the Rescorla–Wagner theory also has a built-in asymmetry between cues and outcomes (see Van Hamme et al., 1993). According to this learning rule, cues compete for the prediction of a common outcome but different outcomes of a single cue do not compete. Thus, the Rescorla–Wagner rule also predicts competition among causes but not among effects *when* the learning situation is set up the right way: when the causes are presented temporally prior to the effects, the Rescorla–Wagner theory reflects the real-world asymmetry between causes and effects. The asymmetry of cues and outcomes has been firmly established in a number of experiments with animals and humans that have demonstrated competition among causes (or cues) but not effects (or outcomes) (Baker & Mazmanian, 1989; Baker, Murphy, & Vallée-Tourangeau, this volume, Ch. 1; Matute, Arcediano, & Miller, 1996, Experiments 1, 2; Van Hamme et al., 1993). All these studies have in common that the causes were presented either prior to or simultaneous with the effects so that the Rescorla–Wagner rule happens to yield the correct predictions.

The Rescorla–Wagner rule, however, makes the wrong predictions when in the learning situation the cues represent effects and the outcomes causes. In these situations this theory predicts competition among the effects but not among the causes, a pattern contrary to that of physical causal relations in the real world. In order to test causal-model theory against standard associationist theories that model learning as the association between cues and outcomes, Waldmann and Holyoak (1992) used a diagnostic learning

situation in which the effects were presented first (as cues) and the feedback about the outcome (the causes of the effects) was given after the participants' diagnostic response. Since the cues and the outcomes were identical in both the predictive and the diagnostic conditions, standard associationist theories predict equal amounts of cue competition in both conditions.

Waldmann and Holyoak's (1992) finding that no cue competition occurred in the diagnostic condition provoked a number of critical responses. Van Hamme et al. (1993) argued that the Rescorla–Wagner rule actually predicts the right pattern when cues are mapped to causes and outcomes to effects. This suggestion, however, faces the problem that it is unclear how the participants of the experiments mastered the diagnostic learning situation in which the effects were presented prior to the causes. It is not clear how an associative network in which the causes represent the input and the effects the output could generate a diagnostic response on the basis of effect cues as the input for their decisions.

Shanks and Lopez (in press) therefore proposed a more complex theory for diagnostic learning. They argued that the participants may run two associative networks in parallel, one that is directed from causes to effects, and one that is directed from effects to causes. The latter network is then responsible for diagnostic learning and the diagnostic inferences from effects to causes. Since in Experiment 1 of Waldmann and Holyoak (1992) participants were requested to give cause–effect ratings in the diagnostic learning conditions, this model appears to explain the observed absence of cue competition.

This theory, however, runs into problems when Experiment 3 of Waldmann and Holyoak (1992) is considered. In this experiment not only the cues and outcomes were held constant; in addition, the test question was identical in both the predictive and the diagnostic learning conditions. Thus, differences in the ratings can be attributed only to the different causal models underlying cues and outcomes. In both learning conditions the participants were asked to rate how "predictive" each individual cue is for the outcome. Therefore, in the diagnostic condition the participants were requested to give a *diagnostic* effect–cause rating. Since in this condition the learning as well as the test question is directed along the effect-to-cause direction, the theory of Shanks and Lopez (in press), along with standard associationist theories, predicts cue competition in the diagnostic condition.

Causal-model theory (Waldmann & Holyoak, 1992) predicts that the participants form causal models in the cause–effect direction but are able to access these representations in both the predictive cause–effect and the diagnostic effect–cause direction. Normatively this implies that the diagnostic inferences should be sensitive to whether a specific effect is caused by only one or by several competing causes. For example, a symptom, such as fever, may be a deterministic effect of a disease. It may nevertheless be a bad diagnostic cue, simply because there are many alternative causes of this symptom. Thus, ratings of whether fever is an effect of the disease should be high, whereas low ratings should be expected when the test question requests an assessment of how predictive it is for the disease.

Waldmann and Holyoak's (1992) Experiments 2 and 3 present a pattern consistent with this prediction. No cue competition was observed in the diagnostic condition with learning materials in which no alternative causes of the state of the effects were given (signal buttons of an alarm; Experiment 3). This finding refutes standard associative theories, including Shanks and Lopez's suggestion. However, reduced ratings for the R cue were obtained when the R cue represented a symptom ("underweight") with many alternative potential causes. Since in Experiment 1 the participants were able to learn that this symptom is an effect of the new disease, the lowering of the ratings with the diagnostic test question in Experiment 2 seems to reflect participants' sensitivity to the difference between predictive (cause–effect) and diagnostic (effect–cause) inferences.

1. *Competition among Causes in Predictive and Diagnostic Learning*

As additional evidence for the assumptions of causal-model theory, Waldmann (1996) designed an experiment ($N = 56$) that directly addresses the question of whether participants are sensitive to the fact that the predictiveness of effect cues is dependent on the presence of alternative causes. A number of critics have pointed out that Waldmann and Holyoak (1992) provided only indirect evidence for this sensitivity since the conclusions were based on a cross-experiment comparison. The more recent experiment may also serve to clarify a misunderstanding. Shanks and Lopez (in press) assert that Waldmann and Holyoak (1992) claim that cue competition in diagnostic learning is dependent on whether the cues are concrete or abstract. What we actually claimed was that people will be sensitive to whether an effect is caused by one or several causes, regardless of whether the effect is concrete or abstract. The more recent experiment used fairly abstract materials and demonstrated sensitivity to the structure of the underlying causal model with identical types of material. Finally, an additional goal of this experiment was to replicate the finding of the absence of a blocking effect in a diagnostic learning task with diagnostic test questions with more abstract kinds of learning materials. Matute et al. (1996, Footnote 1), for example, doubt the validity of the results of Waldmann and Holyoak's (1992) Experiment 3.

In the learning phases in all conditions, participants received information about the presence or absence of different substances in animals' blood,

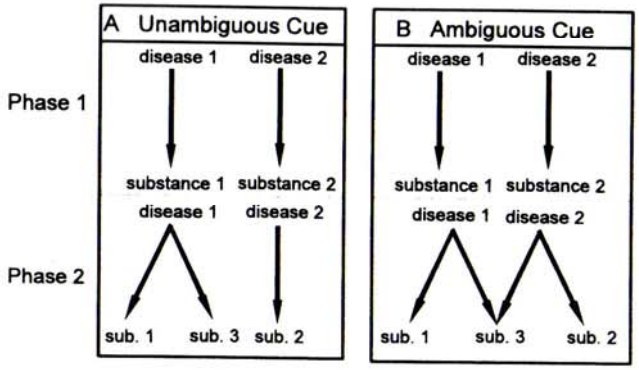

Fig. 4. The two learning phases of the two diagnostic learning conditions, unambiguous (A) and ambiguous (B) cue conditions. The cues (substances) represent potential effects of the causes (diseases) to be diagnosed.

and then they had to judge whether the animal had contracted one of two new blood diseases or not. After each decision feedback was given. The substances were all abstractly numbered and not further characterized (e.g., "Substance 1: Yes; Substance 2: No").

Figure 4 displays the causal structure of the learning domain presented in the *diagnostic* conditions. In these conditions, the substances were characterized as effects of the diseases. Participants were told that new blood diseases had been discovered that produce new types of substances in the blood. In both conditions, the unambiguous and the ambiguous cue condition, participants learned in Phase 1 that substance 1 is caused by disease 1, and substance 2 is caused by disease 2. In Phase 2, however, the two conditions differed. In the *unambiguous cue condition* (A), substance 3 is paired only with substance 1. Participants learned that disease 1 causes substance 1 as well as substance 3.

One of the crucial test questions asked the participants to rate how *predictive* substance 3 is for disease 1. The participants were told that they should imagine being confronted with new animals, and having received information about the presence of only one substance. Their task was to rate how well knowledge about the presence of the respective substance would enable them to predict the existence of the diseases. Associative learning theories, such as the Rescorla–Wagner theory, predict blocking in the unambiguous cue condition (also Shanks & Lopez, in press). The effects are mapped to the input level as the learning and the test questions are directed from effects to causes. Causal-model theory predicts absence

of blocking because the symptom is an effect of the disease and because there are no alternative competing explanations.

In the *ambiguous cue condition* (B), substance 3 is caused by either disease 1 or disease 2. Again, associative theories, including Shanks and Lopez's (in press) proposal, predict complete blocking of the redundant cue. Causal-model theory predicts that participants should be sensitive to the fact that there are multiple explanations for the presence of substance 3. Therefore they should lower their diagnostic ratings in this condition.

The participants rated the predictive cues, substance 1 and 2, high both after Phase 1 as well as after Phase 2. Figure 5 displays the results of Phase 2. The most important result involves the redundant cue in the diagnostic conditions (Fig. 5A). In the unambiguous cue condition, the redundant cue

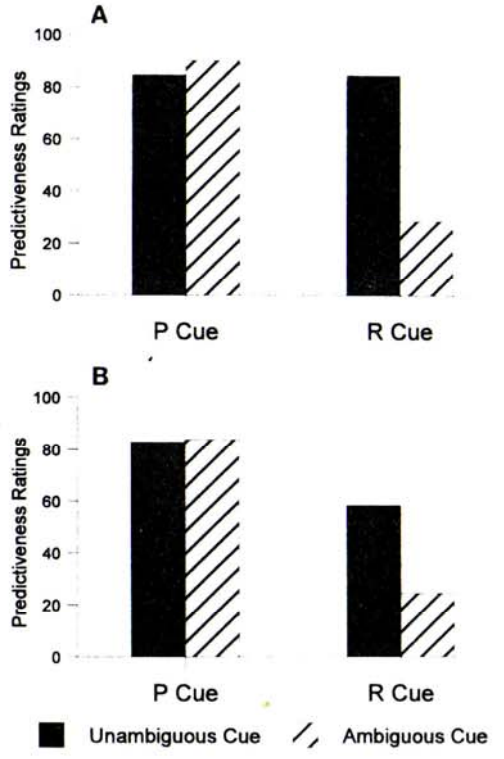

Fig. 5. Mean "predictiveness" ratings from the diagnostic (A) and predictive (B) learning conditions in Phase 2.

yielded high ratings. As in earlier experiments, no sign of blocking can be seen here. As a matter of fact, *all* participants gave identical ratings to the predictive and the redundant cue in this condition. This finding refutes standard associative learning theories. In the ambiguous cue condition the ratings are clearly lowered. The participants were apparently sensitive to the fact that there are competing theories explaining the presence of the redundant effect cue. These two results jointly support causal-model theory.[2]

As a further test of participants' sensitivity to causal directionality, a predictive version of the tank was also investigated. The structure was identical to the one outlined in Fig. 4 except for the fact that the direction of the causal arrows was reversed. In the *predictive conditions* the substances were redefined as potential causes of the new blood diseases. Participants in these conditions were told that some food items appear to contain substances that may cause new blood diseases. The *same* learning exemplars were used as in the diagnostic conditions. The participants received information about the presence or absence of the substances, and then had to judge whether the animal had contracted one of the two diseases or not. Thus, in Phase 1, participants learned that substance 1 causes disease 1, and substance 2 causes disease 2. In Phase 2, substance 3 was redundantly paired only with substance 1 to produce disease 1 (*unambiguous cue condition*), or, in the *ambiguous cue condition*, it was paired with either substance 1 to produce disease 1 or substance 2 to produce disease 2.

Assuming that Phase 1 training was asymptotic, associative learning theories generally predict complete blocking of the redundant cue in both conditions. Causal-model theory also predicts a reduction of the ratings for the redundant cue. However, blocking should be only partial in the unambiguous cue condition. Participants simply do not receive sufficient information to assess the causal status of the redundant cue. Therefore, they should be merely uncertain about whether it is a cause, not certain that it is not a cause. In the ambiguous cue condition, they also receive incomplete information. However, unlike in the unambiguous cue condition, participants see that each disease can also be absent in the presence of the redundant cue. Therefore, in the ambiguous cue condition, they

[2] Van Hamme et al.'s (1993) claim that the Rescorla–Wagner rule explains Waldmann and Holyoak's (1992) results has sometimes been interpreted as the implicit suggestion to generally map causes to the cue level and effects to the output level even when effects are presented first (see, e.g., Matute et al., 1996). It should be noted, however, that this account is also refuted by the results of the experiment, as it would not explain why ratings for the redundant cue were reduced in the ambiguous-cue condition. In both the unambiguous and the ambiguous-cue condition the cause-to-effect contingency of the redundant cue was maximal ($\Delta p = 1$) so that no differences should be expected.

should be more certain that it is not a cause than in the unambiguous cue condition.

Figure 5 (B) displays the means of the Phase 2 ratings in the predictive conditions. For both the ambiguous and the unambiguous cue condition, the redundant cue (R cue) yielded significantly lower ratings than the predictive cue (P cue), which can be interpreted as evidence for blocking in the predictive context. However, blocking was only partial as predicted by causal-model theory. The predicted difference between the ambiguous and unambiguous cue condition was also obtained.[3]

2. *The Role of the Structure of the Causal Model*

Causal-model theory has sometimes been paraphrased as predicting competition among causes but not among effects (Matute et al., 1996). This summarization is incomplete. Waldmann and Holyoak (1992) predicted cue competition in blocking situations when the cues represented potential independent *causes of a common effect*, and the absence of cue competition when the cues represented potential independent *effects of a common cause*. Of course, other causal models are possible and may yield different results. For example, a blocking task could be set up in which the R cue represents a cause of the P cue, which in turn is linked to the effect. This situation instantiates a *causal chain*, and no blocking of the R cue should be expected. Williams, Sagness, and McPhee (1994) have demonstrated that different types of pretraining may indeed influence whether participants view cues as independent or connected (see also Williams, this volume, Ch. 3).

To account for causal chains, the conditional contingency rule (Eq. 2) has to be modified (see Cartwright, 1989; Eells, 1991). Potential cofactors (K) should be kept constant only when they are *not* causal intermediates between the target cause and the target effect. Causal intermediates also screen off the relation between the primary cause and the effect so that

[3] Associative theories may explain a difference between the ambiguous and the unambiguous cue conditions as a consequence of preasymptotic training of the P cue in Phase 1. However, it is unlikely that this account is correct. First, it does not explain the complete absence of blocking in the diagnostic condition. Furthermore, participants had to learn to associate only three simple patterns (either substance 1 or 2 present, or both substances absent) with three responses. This is an extremely easy task and was typically mastered within a couple of trials. Then this account would predict an increase of the ratings of the P cue with increasing training which was not observed (see also Waldmann & Holyoak, 1992). Finally, Waldmann (1996) presents an additional experiment with predictive learning instructions in which the amount of Phase 1 training was varied between either two or ten presentations of each learning exemplar. The Rescorla–Wagner theory predicts an increase of the size of the blocking effect and a decrease of the difference between the ratings of the ambiguous and the unambiguous redundant cue proportional to the amount of Phase 1 training. Causal-model theory predicts no difference. The results clearly supported causal-model theory.

holding them fixed would misrepresent the true causal relations. This is a further example of how prior causal knowledge affects how the statistical relations of the learning input should be processed.

3. *Evidence for Effect Competition?*

Lack of control over the underlying causal structure may lead to apparent refutations of causal-model theory. Shanks and Lopez (in press) present one experiment in which they claim to have found evidence for effect competition (see also Shanks, Lopez, Darby, & Dickinson, this volume, Ch. 7). Shanks and Lopez (in press, Experiment 3) compared two conditions. The "noncontingent" condition presented the following causal structure: cause 1 → AB, cause 1 → B, no cause → C. In the "contingent" condition a different learning structure was used: cause 2 → DE, cause 2 → F, no cause → E. The letters A to F represent effects. These patterns were trained in the diagnostic direction in which the effect cues were presented first.

Standard associative theories that map these effect cues on the input layer predict effect competition. Despite being presented an equal number of times along with the target causes (i.e., diseases), effect A from the noncontingent condition should be rated lower than effect D from the contingent condition. In both conditions, the unconditional contingencies between the cause and the target effect were kept constant so that *prima facie* causal-model theory predicts no difference. Shanks and Lopez discovered a small difference in association ratings between the two conditions, which was interpreted as evidence against causal-model theory.

One problem with this experiment is that the instructions and the cues (symptoms labeled with letters) did not clearly specify the underlying causal model so that it is unclear how the learning input was actually interpreted (see also Waldmann & Holyoak, in press, for a more detailed critique of this study). In this regard, the experiment is similar to previous studies, which, however, never claimed to study causal induction (G. B. Chapman, 1991; Gluck & Bower, 1988a; Shanks, 1991). As pointed out by Waldmann and Holyoak (1992; Footnote 1), not all symptoms of a disease are effects. They may be causes (e.g., puncture wounds indicating blood poisoning), intermediate causes of a causal chain, or part of a complex causal network representing a syndrome.

A second problem is that the learning input points to different underlying causal models. Assuming that the symptoms were actually interpreted as effects as intended by Shanks and Lopez (in press), the noncontingent structure is an instantiation of a simple common-cause model (see Fig. 1A) in which cause 1 deterministically produces symptom B, and weakly

produces symptom A. By contrast, the contingent structure is incompatible with a simple common-cause model. This structure exhibits a situation in which a single cause has disjunctive effects. The disease (i.e., cause 2) causes either the symptom complex DE, or the symptom F, but no other combinations of D, E, and F are ever observed. As a consequence, the initial model would have to be modified to account for the peculiar interaction of the effects. Waldmann et al. (1995) predict greater learning difficulty for the condition with the mismatch between the initially plausible common-cause model and the learning input, which was indeed obtained by Shanks and Lopez (see Waldmann & Holyoak, in press).

Very little is known about the revision processes activated when the initial causal model is incompatible with the learning input (but see Ahn & Mooney, 1995; Waldmann et al., 1995). It is readily apparent, however, that the Rescorla–Wagner model, originally not having been intended to model complex *causal* induction tasks, lacks the flexibility to reconfigure itself in light of evidence incompatible with the implicit causal structure of the learning model.

Esmoris-Arranz et al. (1995) present a study in which they tried to demonstrate effect competition in an animal learning experiment (see also Miller & Matute, this volume, Ch. 4). Assuming that the rats who participated in the experiments actually interpreted the CS as causes and the US as effects, Esmoris-Arranz et al. compared two causal structures. In the experimental condition, the rats learned that a cause A produces an effect S in Phase 1, and in Phase 2 this cause A produces effect S along with a second effect X. In the control condition Phase 2 was identical, but A and S were unpaired during Phase 1. In the test phase rats were presented with the single cues S and X. The most important result involves test cue X that has been paired with A an equal amount of times in the two conditions. Responding to test cue X indicated lower associative weights in the experimental condition than in the control condition. This finding was interpreted by Esmoris-Arranz et al. as evidence for cue competition among the effects S and X.

Again this is a peculiar causal situation when taken at face value. In the experimental condition cause A consistently causes effect S, but it changes its causal power from not producing X during Phase 1 to deterministically producing X in Phase 2. In the control condition, cause A changes from being ineffective to being a deterministic cause of both S and X in Phase 2. It is unclear whether a cause like the one presented in the experimental condition exists in the physical world.

However, even when the unrealistic nature of the presented causal situation is ignored, the results of this experiment do not present unambiguous evidence for effect competition against contingency accounts. K. J. Holyoak

(personal communication) offers a contingency analysis that is consistent with the assumptions of causal-model theory. This analysis assumes that test cue X is implicitly coded as the complex event X and not-S, as the cues S and X have been consistently paired in Phase 2 of the training phase. Thus, in the test situation the rats in the two conditions are actually trying to infer how likely this complex new event (X and not-S) is caused by the unobserved cause A. Although the contingency of A and X is constant across the two conditions, the likelihood that A is producing the absence of S (not-S) seems higher in the control than in the experimental condition. In the control condition, A is paired with the absence of S during Phase 1, whereas A and the absence of S are never combined in the experimental condition. Hence the complex cue X and not-S is less likely to have been caused by A in the experimental than in the control condition, which is in line with the results of the experiment.

Matute et al. (1996) present a different set of experiments in which they tried to provide evidence for effect competition with human participants. They argued (in contrast to the Rescorla–Wagner and many other associative learning theories) that cue competition is a function of the test question that probes the knowledge base, and not a characteristic of the learning rule. In their Experiment 3 they found that the participants tended to rate the relationship of a cause and a specific effect lower when this effect was paired with a stronger as opposed to a weaker second effect (but see Baker & Mazmanian, 1989). The contingency between the cause and the target effect was kept constant across the two conditions. Therefore, this experiment appears to provide *prima facie* evidence for effect competition.

However, this finding crucially depended on the test question. When the participants were asked whether the target effect was an effect of the cause or whether the cause produced this effect, then no effect competition was found (Experiment 2). However, when the test question asked how "indicative" the effect was, then participants tended to give an assessment of the diagnostic validity of the target effect *relative* to the strength of the other collateral effect. It certainly is reasonable that in some circumstances a relative assessment of the diagnostic validity of an effect will be given (as when a physician is about to decide which diagnostic test to conduct). According to causal-model theory, this finding is a further demonstration that the participants are able to flexibly access their knowledge base. People are apparently not only able to access causal knowledge in the predictive and the diagnostic directions, they are also able to compare different causal strengths. An associative learning theory could also account for these data when the assumption is added that in some test situations the responses are based on a choice rule that compares the different associative weights obtained during learning.

Matute et al.'s equivocation of cause and effect competition blurs one of the most fundamental differences between causes and effects, the distinction between *spurious* causes and *collateral* effects. Whether or not a cause is real or spurious may be of the utmost pragmatic importance. It would make little sense to tamper with a barometer when the goal is to influence the weather. By contrast, a redundant, albeit weak effect can be produced regardless of whether there are alternative, maybe stronger effects. The pattern of results Matute et al. (1996) present is consistent with the notion that participants are indeed sensitive to this crucial distinction between spurious or interacting causes and collateral, mutually supporting effects (also Baker & Mazmanian, 1989; Rescorla, 1991, 1993, 1995; Van Hamme et al., 1993; Waldmann & Holyoak, 1992). When the participants were requested to assess *causal* relations, they always proved sensitive to potential competitions among causes but never compared collateral effects, or causes with effects. They were sensitive to the fundamental difference between converging causes and diverging effects. For Matute et al. this pattern of results is simply a result of the semantics of the test question, but this explanation begs the question of *why* participants understand the causal test questions the way they do.

D. Causal Models and the Assessment of Contingencies

Causal directionality is only one aspect of abstract prior causal knowledge influencing the interpretation of the learning input. A further problem of purely bottom-up theories of causality is a consequence of the fact that contingencies between two events may be affected by other causal factors. One solution for this problem, the *conditional contingency* approach, has already been mentioned. According to this theory, contingencies should not be computed over the universal set of events but over subsets of events. However, Cartwright (1983) points out that this method yields correct results only when the subsets are properly selected (see also Cheng, 1993). Conditionalizing on the wrong variables may lead to erroneous contingency estimates. An instance of this problem is known in the philosophical and statistical literature as Simpson's paradox (see Cartwright, 1983; Eells, 1991; Pearl, this volume, Ch. 10; Simpson, 1951).

Simpson's paradox describes the fact that a given contingency between two events that holds in a given population can disappear or even be reversed in all subpopulations, when the population is partitioned in certain ways. Waldmann and Hagmayer (1995) present an experiment that demonstrates Simpson's paradox (see also Spellman, this volume, Ch. 5). Participants were told that importers of tropical fruit are trying to improve the quality of the fruit by irradiating them. However, so far it is unknown

whether the irradiation has a positive, a negative, or no effect on the quality of the fruit. Participants' task in this experiment was to assess the strength of the causal relation between the irradiation of tropical fruit and the quality of fruit using a rating scale ranging from −10 to +10. To assess the efficacy of irradiation, participants received information about the quality of samples of fruit that either had or had not been irradiated. The participants were handed a list, which contained information about 80 samples of fruit. Each sample was represented on one line, and for each sample participants could see whether or not the sample had been irradiated ("yes" or "no"), and whether the quality of this sample was "good" or "bad." In one of the conditions, the condition with the *causally relevant* variable, participants were told that there are two types of fruit, Taringes and Mamones. Additionally it was pointed out that it was expected that irradiation affects these two types of fruit differently. Furthermore, information was added to the list that indicated that one of the two pages showed Taringes, and the other page Mamones.

Table I displays how the cases were distributed. The table displays the proportion of fruit that were of good quality after they were irradiated, and the proportion of fruit that were of good quality without being irradiated. For example, within subgroup A (e.g., Mamones) 36 fruit samples were presented that were irradiated. Forty-four percent of these samples (i.e., 16 out of 36) had good quality after irradiation. As can be seen in Table I, the arrangement of the cases resulted in a reversal of the sign of the contingencies within as opposed to across the grouping variable. Disregarding the grouping variable yields a positive contingency between irradiation and quality of fruit. By contrast, the contingency within each of the subgroups is negative. For half of the participants, the mapping between irradiation and quality of fruit was switched so that these participants saw a symmetric situation with a negative overall contingency, and positive contingencies within the subgroups. The sign of their ratings was reversed in order to make the two subgroups comparable.

TABLE I

CONTINGENCIES AND RELATIVE FREQUENCIES OF FRUIT WITH GOOD QUALITY

	A	B	Total
Irradiation	16/36 (.44)	0/4 (.00)	16/40 (.40)
No irradiation	3/4 (.75)	5/36 (.14)	8/40 (.20)
Contingency	−.31	−.14	+.20

Even though the task for all participants was to assess the overall efficacy of irradiation, it was expected that participants in the condition with the causally relevant grouping variable would assess the causal impact of irradiation separately for each subgroup (Mamones and Taringes), and disregard the total distribution of the cases. Since the contingencies within each subgroup are negative, participants should get the overall impression that irradiation *lowers* the quality of fruit.

This example may lead to the methodological suggestion that it is always a good idea to partition into subsets of events, and compute conditional contingencies in which potential cofactors are kept constant. However, this strategy may also lead to false assessments. The reason why the analysis should be based on the fruit level in the condition with two fruit types is that the fruits are *causally relevant* for the effect under investigation. If, by contrast, it had been shown that the contingencies reverse when the fruits were partitioned on the basis of their position on the test list, this would not count as evidence for a negative causal influence of irradiation. In this situation, one should disregard the groupings, and, based on the total distribution, conclude that irradiation *raises* the quality of fruit. Only partitions by causally relevant variables are relevant for evaluating causal laws (Cartwright, 1983). If causally irrelevant variables also mattered, almost any contingency could be obtained by choosing the right partition of the event space.

In order to test whether participants are sensitive to this crucial distinction between causally relevant and causally irrelevant partitioning variables, a second condition with a *causally irrelevant* variable was included in which participants were told that, due to the large number of tests, the samples of fruit were assigned to different investigators, A and B. Otherwise this condition presented the same learning input, the same assignment of the cases to the two groups, and the same rating instructions as the condition with the causally relevant grouping variable. It was expected that participants in the condition with the causally irrelevant variable would ignore the groups and rely on the total proportions. Thus, they should arrive at the conclusion that irradiation *raises* the quality of fruit. Their ratings should indeed be similar to the ones obtained in an additional control condition in which no grouping information was provided.

Table II shows that participants indeed were sensitive to the distinction between causally relevant and causally irrelevant grouping variables. The ratings in the control condition without a grouping category and in the condition with the irrelevant grouping variable were positive, and statistically indistinguishable from each other. Thus, participants in these two conditions believed that irradiation *raises* the quality of fruit. This finding indicates that the participants based their assessments on the total distribu-

TABLE II

MEAN RATINGS OF THE CAUSAL RELATION BETWEEN IRRADIATION AND QUALITY OF FRUIT

Relevant	Irrelevant	Control
−4.33	5.17	4.75

tion of cases, while disregarding subgroups. By contrast, participants in the condition with the causally relevant grouping variable thought that the cause prevents the effect. These participants concluded that irradiation *lowers* the quality of fruit. Thus, despite the fact that participants in the three conditions received identical learning inputs and identical rating instructions, their assumptions about the causal relevance of an additional grouping variable dramatically influenced their assessment of the relation between a putative cause and an effect.

This example clearly demonstrates that causal induction is crucially dependent on prior causal knowledge. New causal relations may be induced using contingency estimates based on the analysis of the structure of the learning input. However, the contingencies only reflect *causal* relations when the observations are partitioned on the basis of causally relevant rather than irrelevant variables. The causal relevance of *these* partitioning variables has to be established prior to the new induction task. Thus, Simpson's paradox exemplifies the basic assumption of causal-model theory that the processing of the learning input is based on prior assumptions about general properties of the causal situation.

Simpson's paradox is an interesting example of how specific knowledge interacts with abstract causal strategies. It is true that knowledge about the causal relevance of the partitioning variable is domain specific (e.g., the fact that type of fruit is causally relevant). However, unlike in previous research on transfer of specific knowledge (e.g., L. J. Chapman & Chapman, 1967, 1969; Pazzani, 1991), this type of knowledge does not directly bias estimates about the strength of the causal relation between the target cause and the target effect. In order to obtain the correct results, abstract knowledge has to be activated that conditional contingencies based on causally relevant subgroups should be computed. Interestingly, the dramatic reversals obtained in situations exemplifying Simpson's paradox are not due to selective processing of individual cases or knowledge-driven distortions of the contingency estimates. They rather are a natural consequence of unbiased processing of differentially grouped cases.

E. CAUSAL MODELS AND CAUSAL MECHANISMS

The main focus of this article is on the comparison between associative theories of causal induction and causal-model theory. However, since Hume's critical assessment of causal power theories, one of the main debates within the field of causal processing relates to the question of whether causal induction is based on the observation of statistical relations or on the observation of causal mechanisms or continuous causal processes (see Ahn, Kalish, Medin, & Gelman, 1995; Cheng, 1993; Salmon, 1984).

According to causal-model theory, these two positions need not be exclusive. Causal-model theory claims that, in general, statistical input information and prior assumptions about causal processes interact. According to this view, assumptions about causal mechanisms may guide the way the statistical input is processed. Often the mechanisms connecting a cause and an effect are unknown or only partly known. In addition, causal processes cannot be observed directly but have to be inferred on the basis of prior theoretical assumptions and the structure of the observational input (see Cartwright, 1989; Cheng, 1993; Cheng et al., this volume, Ch. 8). Thus, even though causality may not be reducible to mere covariational patterns, statistical relations are a potent way to measure causal processes. Knowledge about causal directionality is one important example of a physical feature that may crucially influence the way the learning input is interpreted (Waldmann & Holyoak, 1992; Waldmann et al., 1995). However, more domain-specific knowledge about causal processes may also play a role.

Waldmann (1991) used a learning paradigm analogous to cue compounding tasks from animal learning paradigms. In one of the experiments (N = 96) the participants learned, for example, that drinking a blue liquid causes a heart rate of +3 in animals. Subsequently, the participants learned that drinking a yellow liquid causes a heart rate of +7. The crucial test question was what would happen when both liquids were mixed and drunk altogether.

In animal learning experiments on cue compounding a typical finding is that two separately trained cues are *additively* integrated when presented in a compound (Couvillon & Bitterman, 1982; Kehoe & Graham, 1988; Weiss, 1972). This finding fits with the *additivity bias* inherent in many associative learning theories (including the Rescorla–Wagner theory).

The participants in the experiment, however, proved sensitive to an additional hint that characterized the causal mechanisms that mediate between causes and effect. In one condition, it was mentioned that the heart rate is affected by the *taste* of the liquids, whereas the other condition characterized the liquids as drugs that could have different *strengths*. Taste is an example of an *intensive* physical quantity, whereas the strength of a

drug represents an *extensive* quantity. Intensive quantities are dependent on proportions and therefore do not necessarily vary with the absolute amount of the substance, whereas extensive quantities vary with amount (see also Reed & Evans, 1987; Wiser & Carey, 1983). Despite the fact that no further domain-related information was given (e.g., about the particular kind of taste), the participants activated general integration rules that were sensitive to this fundamental physical distinction. Generally, significantly more participants computed a weighted *average* of the two causal influences in the taste condition than in the strength condition. Only in the strength condition did an *adding*-type integration turn out to be the dominant rule.

This finding is only one example of how physical knowledge affects the way the learning input is treated. This knowledge may be more concrete than knowledge about causal directionality, but it nevertheless is fairly abstract, as the participants were provided with information about only the general physical characteristics of the causes (intensive vs extensive quantities).

Another example of knowledge-driven processing is prior assumptions about the typical temporal lag between causes and effects. When causes produce their effects with a lag, a naive contingency learning mechanism may never be able to detect the contingency between the distant events. There may also be cases in which a cause produces a dynamic pattern (see Eells, 1991). For example, a drug may be harmful in the short run but cure a disease in the long run.

Gallistel (1990) has pointed out a related problem with associative contingency learning mechanisms. Frequently these theories postulate a trial clock, which determines when a trial starts and when it ends. It can be shown, however, that depending on the size of the trial window, almost every contingency estimate may ensue. A small trial window may, for example, divide a specific CS into three events and represent the following brief temporal lag as the subsequent event (US). This example shows that prior assumptions about what constitutes a potential cause and what constitutes an effect are crucial for obtaining appropriate statistical evidence.

F. Causal Models and the Role of Learning Order

The major goal of the experiments on causal directionality was to demonstrate that the participants of the experiments use their abstract knowledge about the asymmetry of causes and effects when interpreting the learning input. In order to test causal-model theory against associative accounts the experiments kept cues and outcomes constant. This strategy led to a design in which common-effect models were presented in a learning order that

corresponds to predictive reasoning (from causes to effects), whereas the common-cause models were presented in the diagnostic order (from effects to causes). The results clearly show that the participants honored the cause–effect direction regardless of the order in which the constituents of the causal models were presented.

However, it is unlikely that learning order generally has no impact on the mental models that are constructed during learning. The competence participants exhibit in simple causal situations may well break down when confronted with more complex situations. For example, the difficulty of predictive and diagnostic learning probably differs. Bindra, Clarke, and Shultz (1980) have presented experiments that show that children have greater difficulties with diagnostic as compared to predictive inferences. Diagnostic inferences typically involve a retrospective updating or modification of an already-formed mental model. This may be more difficult to accomplish than successively augmenting a causal representation parallel or isomorphic to the unfolding of the causal structure in the observed real world, as happens in predictive learning.

In order to investigate whether learning order affects the kind of information that is acquired, Ulf-D. Reips and I conducted a number of experiments in which the participants learned about fictitious diseases (Waldmann & Reips, 1996). A prototypical example of the causal building blocks used in these experiments is depicted in Figure 4 (Phase 2 of the ambiguous cue condition) in which an M-structure with two diseases and three symptoms is shown. Each disease deterministically causes two symptoms. One of these symptoms is ambiguous, as it is caused by either disease. The other symptom is unique. It is produced by only one of the diseases.

Unlike in the previous reported experiments in which the direction of the causal arrow was varied across conditions, the causal structures were kept constant across the learning conditions in this set of studies. Thus, in all conditions the symptoms represented effects and the diseases causes.

Two basic learning conditions were compared. The participants acquired the information about the diseases either in the *predictive* direction, or in the *diagnostic* direction. In the predictive learning condition the participants were presented with information about the disease of a patient, and they had to learn to predict what (two) symptoms this patient probably would exhibit. Each patient was affected by only one of the diseases. In the diagnostic learning condition, the participants received information about the (two) symptoms of each patient first, and had to learn to diagnose the patient's disease. The crucial question was whether these two modes of acquiring knowledge about identical causal structures would lead to different representations. Besides its theoretical relevance, this question is also practically significant. Medical knowledge, for example, is typically pre-

sented in textbooks in the predictive direction regardless of the fact that later this knowledge frequently has to be used in the diagnostic direction.

In one set of experiments we varied the *base rates* of the diseases. One of the diseases of the M-structure was presented three times as often as the other (see also Medin & Edelson, 1988). The participants were trained in either the predictive or the diagnostic direction and then all participants were asked questions about the diagnostic validity of each individual symptom. The most important question involved the ambiguous symptom. Since it is caused by either disease, it would be appropriate to choose the more frequent disease when only information about the presence of this symptom is available.

We predicted that base rate appreciation should be higher after diagnostic than after predictive training. According to contingency theories, predictive learning involves estimating the probability of the effects conditional on the presence and on the absence of the causes. As long as the relevant conditional probabilities are kept constant this estimate is indpendent of whether the causes are frequent or rare (see also Cheng & Novick, 1991). Thus, it was expected that these frequencies should be disregarded in predictive learning. Causal-model theory additionally predicts the asymmetry of learning conditions as a consequence of knowledge about causal directionality. Causes are events that are often actively set in order to achieve effects. For example, when planning psychological experiments it is recommended to establish equal cell sizes. Thus, the observed frequency of the causes is often not representative of the natural frequency of these events in the world, and should therefore not be accepted at face value (see also Gigerenzer, Hell, & Blank, 1988). Even though the cover stories used in our experiments emphasize that participants are going to see unselected samples of patients, there may be a tendency to later disregard this information. This tendency may be stronger when the learning task is complex so that it becomes more difficult to keep in mind additional information that is relevant only for the transfer task. By contrast, diagnostic inferences are based on the observation of effects that cannot be directly manipulated. Thus, the frequency of the causes responsible for the observed effects is generally more representative. Furthermore, diagnostic learning involves the appreciation of base rates. Since the participants in the diagnostic training condition received direct feedback about the disease causing the observed symptoms, the importance of the use of base rates may be experienced more directly (see also Gluck & Bower, 1988a; Griffin & Tversky, 1992; Holyoak & Spellman, 1993; Klayman & Brown, 1993; Koehler, in press; Medin & Edelson, 1988).

In one of the experiments ($N = 32$), six diseases and nine symptoms (three M-structures) were presented either in the predictive or in the diagnostic

learning direction. Within each M-structure one disease was presented three times as often as the other. After the learning phase, the participants were told that they should assume that they were confronted with new patients, and that they knew about the presence of only one of the symptoms of these patients. Their task was to rate the probability of the diseases on a scale from 0 ("very unlikely") to 100 ("very likely"). The most important results involve the ratings of the ambiguous symptoms. Figure 6A displays the mean ratings of the probability that the frequent and the rare diseases were present (collapsed over the three ambiguous symptoms). The results exhibit an interaction. After diagnostic training, the participants proved sensitive to the different base rates of the diseases. They gave higher ratings to the more frequent diseases than to the rare diseases. Base rates were, however, neglected after predictive training. There were no significant

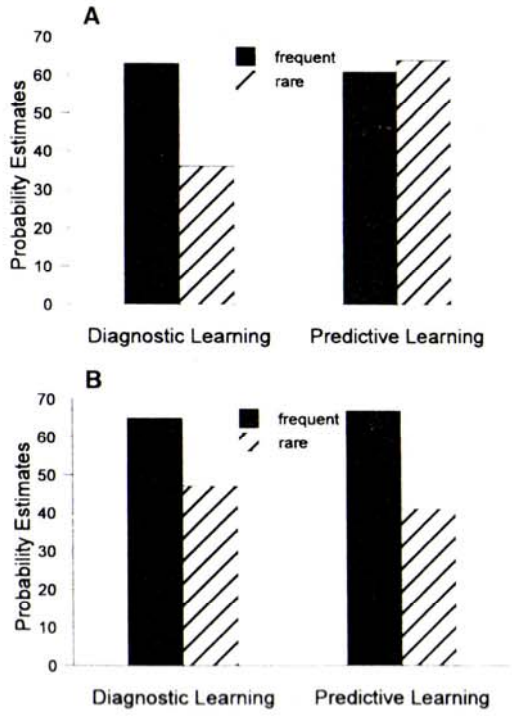

Fig. 6. Mean ratings of the probabilities of the rare and the frequent diseases conditional on the ambiguous symptom after the complex (A; three M-structures) and the simple (B; one M-structure) learning tasks.

differences between the ratings in this training condition. This neglect was not the result of a failure to encode frequencies, as the participants in both conditions turned out to be fairly good at remembering the different frequencies of the disease. Thus, the learning order seemed to affect the tendency to *use* base rates rather than the encoding.

In a second experiment, a situation with only one M-structure (two diseases, three symptoms) was presented ($N = 24$). As can be seen in Figure 6B, this reduction in complexity led to base rate appreciation after both learning situations. Thus, the results in the first experiment were not caused by a general deficit of participants' competence. It rather reflected a performance factor. The competence could be exhibited only in relatively simple learning situations. When the complexity of the learning structure was increased, additional, performance factors came into play, which led to a tendency to neglect base rate information after predictive training. The pattern of results exemplified in these two experiments has been replicated in a number of additional experiments. These findings clearly show that, at least with more complex causal situations, learning order may affect the way mental models are formed and accessed.

IV. Conclusion

The comparison between causal-model theory and associative accounts of causal induction highlighted a number of important differences between these two approaches. Causal-model theory postulates a rigorous separation between the learning input and mental representations. This characteristic allows for the flexible assignment of the learning input to elements of the resulting mental models. By contrast, most associative learning theories (e.g., the Rescorla–Wagner theory) work in the tradition of stimulus–response theories in which learning cues play the double causal role of representing events and eliciting responses. It has been shown that this inflexibility may lead to clear misrepresentations of objective causal relations. Most saliently, associative theories that code the learning cues as CS and the outcomes as US are unable to capture the structural characteristics of diagnostic learning situations in which effects are presented as cues. The Rescorla–Wagner theory correctly captures the asymmetry between causes and effects only when the learning situation is fortuitously presented in a way that corresponds to the implicit structural characteristics of this theory.

A second major tenet of causal-model theory postulates the necessity of an interaction between top-down assumptions and the processing of the learning input. Here, causal-model theory represents a reconciliation between theories focusing on statistical covariation learning and theories

focusing on causal, mechanical processes. Causal-model theory is consistent with Cartwright's (1989) philosophical analyses of causality. Cartwright views causes as entities that embody an intrinsic dispositional *capacity* to produce effects. For example, smoking has the capacity to produce lung cancer. Due to additional causal factors, this capacity may not materialize in all contexts but it may still manifest itself in probabilistic relationships. Thus, covariation is one of the most potent ways to *measure* causal capacities. Like other measuring instruments it needs to be read properly. Covariation does not directly define causality. In this article a number of studies have been presented that demonstrated how identical learning inputs may be processed differently depending on participants' background assumptions about the causal processes to be observed.

Causal directionality is one of the most important features of causal relations that determine the way statistical relations are interpreted. It is a physical fact that multiple causes of a common effect potentially interact, whereas multiple effects of a common cause are rendered conditionally independent when the common cause is held constant (Reichenbach, 1956). Knowing that causes enable us to produce effects, and that redundant causes as opposed to redundant effects may be spurious, is highly relevant for planning our actions. Associative theories imply that participants are unaware of these fundamental distinctions. However, a number of studies have been presented in this chapter that show that participants are indeed sensitive to causal directionality and the asymmetry of causes and effects (see also Waldmann & Holyoak, 1990, 1992; Waldmann et al., 1995).

Assumptions about causal directionality are only one example of how prior knowledge may guide the induction process. Taking into account alternative causal factors is another important method of measuring causal capacities. In many situations, simple unconditional contingencies do not correctly reflect the underlying causal relations. When alternative causal factors are present, conditional contingencies should be computed that hold these factors constant (Cartwright, 1989; Cheng, 1993; Cheng & Novick, 1992; Melz et al., 1993). However, even this recommendation leads to correct results only when the right background conditions hold. A cofactor should be taken into account only when it is expected to be *causally relevant* (Waldmann & Hagmayer, 1995). Furthermore, causal factors should not be used as conditioning variables when they constitute intermediates in a causal chain linking the target cause and the target effect (Cartwright, 1989), or when they represent collateral side effects (Eells, 1991). In these cases, holding the cofactors fixed distorts the statistical relations between the target cause and the target effect, and prevents the causal factor from displaying its causal significance in the form of the relevant conditional probabilities. These are examples of how prior assumptions about the causal

model underlying the observed events may dramatically alter the way statistical information should be processed. Other examples of effects of prior knowledge include assumptions about the integration of causal influences (Waldmann, 1991), about the temporal lag between causes and effects (Anderson, 1990), about the mathematical function relating continuous causes and effects (Zelazo & Shultz, 1989), and about the segmentation of the event stream into potential causes and effects (Gallistel, 1990). Without prior knowledge that is already available at the outset of the induction process new causal knowledge cannot properly be acquired.

Acknowledgments

I would like to thank P. Cheng, D. Medin, A. Merin, U.-D. Reips, and D. Shanks for helpful comments. I particularly thank Keith Holyoak for many stimulating discussions.

References

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

Ahn, W., & Mooney, R. J. (1995). Biases in refinement of existing knowledge. In J. D. Moore & J. F. Lehman, (Eds.), *Proceedings of the Seventeenth Annual Conference of the Cognitive Science Society* (pp. 437–442). Hillsdale, NJ: Erlbaum.

Alloy, L. B., & Tabachnik, N. (1984). Assessment of covariation by humans and animals: The joint influence of prior expectations and current situational information. *Psychological Review, 91,* 112–149.

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

Baker, A. G., & Mazmanian, D. (1989). Selective associations in causality judgments II: A strong relationship may facilitate judgments of a weaker one. In *Proceedings of the Eleventh Annual Conference of the Cognitive Science Society* (pp. 538–545). Hillsdale, NJ: Erlbaum.

Bindra, D., Clarke, K. A., & Shultz, T. R. (1980). Understanding predictive relations of necessity and sufficiency in formally equivalent "causal" and "logical" problems. *Journal of Experimental Psychology: General, 109,* 422–443.

Bromberger, S. (1966). Why-questions. In R. Colodny (Ed.), *Mind and cosmos* (pp. 86–114). Pittsburgh: University of Pittsburgh Press.

Cartwright, N. (1983). *How the laws of physics lie. Essay 1.* Oxford: Clarendon Press.

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

Chapman, G. B. (1991). Trial order affects cue interaction in contingency judgment. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 17,* 837–854.

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 causal facts: Pressing the limits of statistical relevance. In D. L. Medin (Ed.), *The psychology of learning and motivation* (Vol. 30, pp. 215–264). San Diego, CA: Academic Press.

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.

Choi, S., McDaniel, M. A., & Busemeyer, J. R. (1993). Incorporating prior biases in network models of conceptual rule learning. *Memory & Cognition, 21,* 413–423.

Couvillon, P. A., & Bitterman, M. E. (1982). Compound conditioning in honeybees. *Journal of Comparative and Physiological Psychology, 96,* 192–199.

Dawes, R. M. (1988). *Rational choice in an uncertain world.* San Diego, CA: Harcourt Brace Jovanovich.

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

Eddy, D. M. (1982). Probabilistic reasoning in clinical medicine: Problems and opportunities. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 249–267). New York: Cambridge University Press.

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

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

Esmoris-Arranz, F. J., Miller, R. R., & Matute, H. (1995). *Blocking of antecedent and subsequent events: Implications for cue competition in causality judgment.* Manuscript submitted for publication.

Fair, D. (1979). Causation and the flow of energy. *Erkenntnis, 14,* 219–250.

Gallistel, C. R. (1990). *The organization of learning.* Cambridge, MA: MIT Press.

Gigerenzer, G., Hell, W., & Blank, H. (1988). Presentation and content: The use of base rates as a continuous variable. *Journal of Experimental Psychology: Human Perception and Performance, 14,* 513–525.

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

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

Griffin, D., & Tversky, A. (1992). The weighing of evidence and the determinants of confidence. *Cognitive Psychology, 24,* 411–435.

Hempel, C. G. (1965). *Aspects of scientific explanation.* New York: Free Press.

Holyoak, K. J., & Spellman, B. A. (1993). Thinking. *Annual Review of Psychology, 44,* 265–315.

Hume, D. (1977). *An enquiry concerning human understanding.* Indianapolis, IN: Hackett Publishing Company. (Original work published 1748).

Hume, D. (1978). *A treatise of human nature.* Oxford: Clarendon Press. (Original work published 1739).

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

Kamin, L. J. (1969). Predictability, surprise, attention, and conditioning. In B. A. Campbell & R. M. Church (Eds.), *Punishment and aversive behavior* (pp. 276–296). New York: Appleton-Century-Crofts.

Kant, I. (1950). *Critique of pure reason* (N. K. Smith, Trans.). London: Macmillan. (Original work published 1781).

Kehoe, E. J., & Graham, P. (1988). Summation and configuration: Stimulus compounding and negative patterning in the rabbit. *Journal of Experimental Psychology: Animal Behavior Processes, 14,* 320–333.

Kitcher, P. (1989). Explanatory unification and the causal structure of the world. In P. Kitcher & W. C. Salmon (Eds.), *Minnesota studies in the philosophy of science* (Vol. 13, pp. 410–505). Minneapolis: University of Minnesota Press.

Klayman, J., & Brown, K. (1993). Debias the environment instead of the judge: An alternative approach to reducing error in diagnostic (and other) judgment. *Cognition, 49,* 97–122.

Koehler, J. J. (in press). The base rate fallacy reconsidered: Descriptive, normative and methodological challenges. *Behavioral and Brain Sciences.*

Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review, 99,* 22–44.

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

Mackintosh, N. J. (1983). *Conditioning and associative learning.* Oxford: Clarendon Press.

Matute, H., Arcediano F., & Miller, R. R. (1996). Test question modulates cue competition between causes and between effects. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22,* 182–196.

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

Melz, E. R., Cheng, P. W., Holyoak, K. J., & Waldmann, M. R. (1993). Cue competition in human categorization: Contingency or the Rescorla-Wagner learning rule? Comments on Shanks (1991). *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19,* 1398–1410.

Minsky, M., & Papert, S. (1969). *Perceptrons: An introduction to computational geometry.* Cambridge, MA: MIT Press.

Murphy, G. L., & Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychological Review, 92,* 289–316.

Pavlov, I. P. (1927). *Conditioned reflexes.* London: Oxford University Press.

Pazzani, M. J. (1991). Influence of prior knowledge on concept acquisition: Experimental and computational results. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 17,* 416–432.

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

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

Reed, S. K., & Evans, A. C. (1987). Learning functional relations: A theoretical and instructional analysis. *Journal of Experimental Psychology: General, 116,* 106–118.

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

Rescorla, R. A. (1973). Evidence for the "unique stimulus" account of configural conditioning. *Journal of Comparative and Physiological Psychology, 85,* 331–338.

Rescorla, R. A. (1991). Associations of multiple outcomes with an instrumental response. *Journal of Experimental Psychology: Animal Behavior Processes, 17,* 465–474.

Rescorla, R. A. (1993). Preservation of response-outcome associations through extinction. *Animal Learning and Behavior, 21,* 238–245.

Rescorla, R. A. (1995). Full preservation of a response-outcome association through training with a second outcome. *The Quarterly Journal of Experimental Psychology, 48B,* 252–261.

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

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning internal representations by error propagation. In D. E. Rumelhart, J. L. McClelland, & The PDP Research Group (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition* (Vol. 1, pp. 318–362). Cambridge, MA: MIT Press.

Salmon, W. C. (1971). *Statistical explanation and statistical relevance.* Pittsburgh: University of Pittsburgh Press.

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

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

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

Shanks, D. R. (1991). Categorization by a connectionist network. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 17,* 433–443.

Shanks, D. R. (1993). Human instrumental learning: A critical review of data and theory. *British Journal of Psychology, 84,* 319–354.

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

Shanks, D. R., & Lopez, F. J. (in press). Causal order does not affect cue selection in human associative learning. *Memory & Cognition.*

Simpson, E. H. (1951). The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society, Series, B, 13,* 238–241.

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

Thorndike, E. L. (1911). *Animal intelligence: Experimental studies.* New York: Macmillan.

Tolman, E. C., & Brunswik, E. (1935). The organism and the causal texture of the environment. *Psychological Review, 42,* 43–77.

Tversky, A., & Kahneman, D. (1980). Causal schemas in judgments under uncertainty. In M. Fishbein (Ed.), *Progress in social psychology* (pp. 49–72). Hillsdale, NJ: Erlbaum.

Van Hamme, L. J., Kao, S. F., & Wasserman, E. A. (1993). Judging interval relations: From cause to effect and from effect to cause. *Memory & Cognition, 21,* 802–808.

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

Waldmann, M. R. (1991, November 23–25). *Cue-compounding versus cue-decompounding of complex causes.* Paper presented at the 32nd annual meeting of the Psychonomic Society, San Francisco.

Waldmann, M. R. (1996). *Competition among causes in predictive and diagnostic learning.* Manuscript in preparation.

Waldmann, M. R., & Hagmayer, Y. (1995). When a cause simultaneously produces and prevents an effect. In J. D. Moore & J. F. Lehman (Eds.), *Proceedings of the Seventeenth Annual Conference of the Cognitive Science Society* (pp. 425–430). Hillsdale, NJ: Erlbaum.

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

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

Waldmann, M. R., & Holyoak, K. J. (in press). Determining whether causal order affects cue selection in human contingency learning: Comments on Shanks and Lopez (in press). *Memory & Cognition.*

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., & Reips, U.-D. (1996). *Base rate appreciation after predictive and diagnostic learning.* Manuscript in preparation.

Wasserman, E. A. (1990). Detecting response-outcome relations: Toward an understanding of the causal texture of the environment. In G. H. Bower (Ed.), *The psychology of learning and motivation* (Vol. 26, pp. 27–82). San Diego, CA: Academic Press.

Wasserman, E. A. (1993). Comparative cognition: Beginning the second century of the study of animal intelligence. *Psychological Bulletin, 113,* 211–228.

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

Weiss, S. J. (1972). Stimulus compounding in free-operant and classical conditioning: A review and analysis. *Psychological Bulletin, 78,* 189–208.

Williams, D. A., Sagness, K. E., & McPhee, J. E. (1994). Configural and elemental strategies in predictive learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20,* 694–709.

Wiser, M., & Carey, S. (1983). When heat and temperature were one. In D. Gentner & A. Stevens (Eds.), *Mental models* (pp. 267–297). Hillsdale, NJ: Erlbaum.

Wisniewski, E. J., & Medin, D. L. (1994). On the interaction of theory and data in concept learning. *Cognitive Science, 18,* 221–281.

Young, M. E. (1995). On the origin of personal causal theories. *Psychonomic Bulletin & Review, 2,* 83–104.

Zelazo, P. D., & Shultz, T. R. (1989). Concepts of potency and resistance in causal prediction. *Child Development, 60,* 1307–1315.