(1995). In H.L. Roitblat + J.-A. Meyer (Eds.), Comparative approaches to cognitive science (pp. 271-302). Cambridge, MA: MIT Press.

12 Complex Adaptive Systems as Intuitive Statisticians: Causality, Contingency, and Prediction

Patricia W. Cheng and Keith J. Holyoak

INTRODUCTION

Any complex adaptive system—whether a human, some other type of animal, or an intelligent machine—that operates in a realistic environment must be able to induce causal connections among events. Causal knowledge is required to predict future states of the environment and consequences of the system's own actions. In addition, causal knowledge can potentially be used to generate and evaluate explanations of why significant events occur (or fail or occur). Part of this knowledge is based on statistical regularity among events.

At least for the past quarter century, many psychologists have seriously considered the possibility that untutored humans as well as other animals are capable of acquiring and using statistical knowledge about the structure of the environment. Peterson and Beach (1967) called people "intuitive statisticians," and Kelley (1967) proposed that people are "intuitive scientists." In the context of experimental paradigms investigating classical conditioning, other theorists have suggested that lower animals operate as intuitive statisticians (e.g., Gallistel 1990; Miller and Schachtman 1985). Although there has in fact been broad agreement that various forms of causal induction depend on the implicit computation of statistical information, the question of precisely what is computed has yet to be resolved. In the field of animal conditioning, as well as in human categorization and causal induction, various theorists have proposed that animals perform some implicit computation of statistical contingency, the difference between the proportion of events for which an effect occurs when a factor is present and that proportion when it is absent.

In all these fields, the contingency approach has been contrasted with the associationist approach exemplified by the connectionist learning rule incorporated in the Rescorla and Wagner (R-W) model of conditioning (Rescorla and Wagner 1972). The R-W model is directly related to a number of issues that lie at the very core of cognitive science. The R-W model was originally proposed as a model of classical conditioning in animals; however, a number of researchers have extended the model to

account for apparently higher-order learning in humans, such as categorization and causal induction. The R-W learning rule is equivalent to the least mean squares (LMS) learning rule that is commonly used to adjust the weights on links in connectionist networks (Widrow and Hoff 1960; see also Sutton and Barto 1981). Gluck and Bower (1988), for example, have applied an adaptive connectionist network using the LMS rule to model data on human categorization (also Estes et al. 1989). Similarly, Shanks (1991) applied a connectionist implementation of the R-W model to attempt to account for the effects of cue competition in a task involving classification of diseases on the basis of symptoms (see also Chapman and Robbins 1990; Wasserman 1990). Because of its apparent simplicity and evident generality, the R-W model remains highly influential as an approach to inductive learning in adaptive systems.

Some theorists have argued that the R-W model can account for phenomena involving cue competition and other cue interactions that cannot be explained by contingency. We shall argue, however, that these advantages claimed for the R-W model over contingency theory disappear when the concept of contingency is suitably generalized along lines suggested by a number of philosophers and psychologists. In fact, the R-W model can itself be analyzed as a mechanism that computes contingency under a certain restricted condition that we will discuss later. For such cases, the R-W model is successful in predicting cue competition, although even here its successes are qualified for domains in which the adaptive system operates on representations coded in terms of the probabilities of events, for which the additivity assumption underlying the model is inappropriate. Outside of cases that satisfy the restricted condition, the R-W model does not compute contingency, and in such situations the model appears to be empirically inadequate. In contrast, a generalized contingency theory can explain a number of the phenomena that contradict predictions of the R-W model.

In this chapter we present a contingency analysis of the successes and failures of the R-W model. Our theoretical analysis may provide a framework for understanding the weakness of an important associationist model. We hope that it will guide future research concerning how statistical regularity is computed by adaptive systems to infer the causal structure of their environments in the course of learning, on the basis of which predictions are made.

WHAT IS COMPUTED IN ASSESSING REGULARITY?

The Probabilistic Contrast Model

It has long been argued that contingency is a component of the normative criterion for inferring a causal link between a factor and an effect (e.g., Kelley 1967; Rescorla 1968; Salmon 1980). Cheng and Novick (1990) proposed an extended version of contingency theory, which they termed the *probabilistic contrast model* (PCM), as a descriptive account of the use of statistical regularity in human causal induction. The model, which applies to events describable by discrete variables, assumes that one of the initial criteria for identifying potential causes is perceived temporal priority (i.e., causes must be perceived to precede their effects). The model assumes that potential causes are then evaluated by contrasts computed over a focal set. (We shall use the terms *contrast* and *contingency* interchangably.) The focal set for a contrast is a contextually determined set of events that the reasoner selects to use as input to the computation of that contrast. It is often not the universal set of events, contrary to what has been assumed by previous contingency theories in psychology. Consider the set of events selected for inferring what causes a forest fire. Reasoner will normally restrict their focal set to terrestrial events, in which oxygen is always present, and will not consider events that occur in oxygen-free outer space.

Using the events in the focal set, a *main effect* contrast specifying a potential cause i is defined as

$$\Delta p_i = p_i - p_{\bar{i}}, \tag{1}$$

where p_i is the proportion of events for which the effect occurs when factor i is present, and $p_{\bar{i}}$ is the proportion of events for which the effect occurs when factor i is absent. (The proportions are estimates of the corresponding conditional probabilities.) If Δp_i is noticeably different from 0, i is perceived as a cause. Note that, if a factor i is constantly present within the focal set, the second term in the contrast, $p_{\bar{i}}$, cannot be calculated. Thus in our forest fire example it will be impossible to compute a contrast for oxygen, since this factor is never absent within the focal set of terrestrial events; as a result, oxygen will not be considered a cause of the fire (even though people would agree, if probed, that the presence of oxygen was in fact necessary for the fire to have occurred).

Contrasts can be either positive, in which case the cause is *excitatory*, or they can be negative, in which case the cause is *inhibitory*. For example, smoking presumably has a positive contrast with respect to lung cancer, and hence can be viewed an an excitatory cause of the disease; whereas exercise has a negative contrast with respect to heart disease, and hence can be viewed as an inhibitory cause of the disease. Confidence in the assessment of a contrast is presumed to increase monotonically with the number of cases observed.

Cheng and Novick (1995) show that, for situations in which alternative causes occur and act independently of i, a positive main effect contrast for i gives an estimate of the causal power of i, as represented by the probability with which i produces the effect. This estimate is unbiased when alternative causes are absent within the focal set and/or the alternative factors present do not produce the effect. To the extent that these

conditions are violated, the contrast for i tends to be an underestimate of the power of i. In the extreme case in which some alternative cause is always present within the focal set and it always produces the effect, the contrast for i, which is zero, is uninterpretable.

The above derivation also shows that, for situations in which alternative causes do not occur independently of i, the main effect contrast for i is confounded by the influence of these causes and is not interpretable as an estimate of the power of i. To eliminate this confounding, it is therefore important to compute what we will call *conditional contrast*, the contrast for the candidate factor that is conditional on holding constant the status of one (or more) other factors. A number of philosophers have proposed conditional contrasts as a criterion for inferring causality (see Cartwright 1983, 1989; Reichenbach 1956; Salmon 1980; Suppes 1970).

Main effect contrasts assess the causal status of each factor considered individually. However, it is also possible for combinations of factors to influence the effect in ways that could not be predicted by the independent influences of the individual factors. Such situations involve interactions between factors, which can be assessed by means of a generalization of main effect contrasts (Cheng and Novick 1995). For example, a two-way interaction contrast specifying the conjunction of potential causal factors i and j is defined as

$$\Delta p_{ij} = (p_{ij} - p_{\bar{i}j}) - (p_{i\bar{j}} - p_{\bar{i}\bar{j}}) + (p_{ij} \cdot p_{\bar{i}\bar{j}}) - (p_{i\bar{j}} \cdot p_{\bar{i}j}) \quad (2)$$

where p, as before, denotes the observed proportion of cases in which the effect occurs when a potential contributing factor is either present or absent, as denoted by its subscripts. If Δp_{ij} is noticeably greater than zero, then i and j combine to produce the effect.[1] A two-way interaction contrast is thus based on a difference of differences—here, the contrast for i when j is present minus the contrast for i when j is absent—with the nonadditivity of probabilities taken into account by the product terms in equation 2. Suppose, for example, that there are two drugs, A and B, which are safe when taken individually but usually fatal when taken together. The contrast for drug A with respect to death will therefore be high when B is also present, but zero when B is absent. Both product terms will also be zero. Accordingly, the interaction contrast (the difference between the above two contrasts corrected by the product terms) will be high.

Notice that each of the two constituent contrasts in an interaction contrast is a conditional contrast. Also notice that conditional contrasts can be described in terms of variations of the focal set. We could say that, in the focal set of events in which drug B is administered, the contrast for drug A is high whereas, in the focal set of events in which drug B is not administered, the contrast for drug A is zero. Furthermore, both of these conditional contrasts will differ from the unconditional contrast for A. This unconditional contrast is equivalent to the main effect contrast for drug A over the focal set of all events involving the presence or absence of drug B.

Cheng and Novick (1990, 1991, 1992) have provided support for contrasts computed over an accurately identified focal set as a descriptive model of human causal inference. The model successfully predicts simple and conjunctive causal attributions and explains a number of empirical phenomena involving human causal attributions that had previously been considered biases. To illustrate the role of focal sets, consider the psychological distinction between causes and enabling conditions. In our example about what causes a forest fire, people might consider a lightning strike as the cause, but they will view the presence of oxygen as merely an enabling condition. Although oxygen is necessary for the fire, it is constant in the relevant focal set so that a contrast cannot be computed. Notice that, in a different context, which evokes a focal set within which the presence of oxygen may vary (for example, a special laboratory intended to be oxygen-free), oxygen will be considered the cause of a fire that breaks out when oxygen leaks into that environment. The assessment of causation thus depends on pragmatic contextual influences. In terms of PCM, a potential causal factor that covaries with the effect (i.e., has a noticeable contrast with respect to the effect) within the contextually determined focal set (e.g., lightning with respect to forest fires in the context of a forest) will be viewed as a cause, whereas a factor that is constant within that focal set (e.g., oxygen in a forest), but is known to covary with the effect in some other focal set (e.g., oxygen covaries with fire in special environments in which the occurrence of oxygen varies), is viewed as an enabling condition. As will be elaborated later, an enabling condition can be distinguished from an alternative cause that happens to be constantly present in the current focal set. We will return to the challenge that the distinction between causes and enabling conditions poses for R-W.

The most central claim of contingency theory, which is reflected in PCM, is that causal attributions depend not only on the probability of the effect given the presence of a cue, but also on the probability of the effect given the absence of the cue. In other words, a cue is viewed as causal only if its presence makes a difference to the probability of the effect. However, theorists have often resisted the notion that humans and other animals implicitly tally information about what happens in the absence of a potential cause. In particular, associationist models of animal conditioning have eschewed any direct representation of cause-absent information. One apparent reason for the reluctance to posit representations of cause-absent information is that any event could potentially be defined in terms of an indefinitely large number of absent factors. It would indeed be bizarre to suppose, for example, that your understanding of this passage might be caused by the (presumed) absence of ravens in the room in which you now sit. The generalized contingency model addresses

this problem by restricting the initial tabulation of cause-absent information to those factors that are plausible causes according to prior knowledge or according to observed pairing with the effect. The challenge for associationist models has been to account for apparent influences of contingency on learning without introducing representations of the absence of potential causal factors. As we will see, the empirical successes and failures of the R-W model can be differentiated by when it succeeds or fails to implicitly tally cause-absent information.

The Rescorla-Wagner Model

The most influential associationist theory of conditioning over the past two decades has been the R-W model. Interestingly, Rescorla (1968) was a harbinger of the importance of contingency in classical conditioning, which he demonstrated with elegant experiments showing that conditioning depends on events that occur in the absence as well as the presence of a cue. Nonetheless, he then went on to develop an associationist model that avoided postulation of representations of the absences of cues. The R-W model represents the learning of an association between cue i (e.g., a tone that is present in the current event) and outcome j (e.g., shock) by a change in the strength of a link between two elemental units, one representing cue i, and the other representing outcome j. (Cue i and outcome j are traditionally termed the *conditioned stimulus* and the *unconditioned stimulus*, respectively.) For any cue i that is present during the event, strength is revised according to the rule

$$\Delta V_{ij} = \alpha_i \beta_j \left(\lambda_j - \sum_{k=1}^{n} V_{kj} \right), \tag{3}$$

where ΔV_{ij} is the change in associative strength between cue unit i and outcome unit j as a result of the current event, α_i and β_j are rate parameters that depend on the salience of i and j, respectively, and λ_j is the desired output corresponding to the actual outcome. Typically, if the outcome is present, λ_j is defined as 1; if the outcome is absent, this value is defined as 0.

$$\sum_{k=1}^{n} V_{kj},$$

defined as the sum of the current strengths of links to unit j from all units representing the n cues present in that event, is the actual output of the network predicting the outcome. If cue i is not present during the event, the associative strength of its cue unit remains unchanged. (The absence of a cue is not represented by any unit.) Learning continues until there is no discrepancy between the desired and actual outputs (averaged over a number of trials). In addition to the particular stimuli present (e.g., a tone), the cues are assumed to include one that represents a context present in every event (e.g., the conditioning cage). In causal terms, each cue i is a potential cause, and j is the effect. The strengths that are updated according to equation 3 are equivalent to weights on the links in a two-layered connectionist network, with the predicting cues represented on the input layer and the predicted outcome on the output layer.

A major attraction of R-W is its ability to explain the effects of interaction between cues. For example, it predicts the phenomenon of *blocking* (e.g., Kamin 1969; Rescorla 1981). Let P be a previously trained predictive cue (i.e., the presence of P has been paired with the outcome, and the absence of P has been paired with the absence of the outcome). Consider the situation in which a novel cue, R, in combination with P, is paired with the outcome. It has been shown that, despite the positive unconditional contrast for R, the learning of this cue is blocked if it is presented only in combination with P. According to R-W, learning occurs only when there is some discrepancy between the predicted and actual outcomes. Because a predictive cue fully predicts the outcome as a consequence of prior pairings, no conditioning would accrue to R.

Rescorla (1968) demonstrated that no conditioning accrues at asymptote to a cue if the effect occurs equally often in its absence as in its presence. The R-W model explains this effect of contingency by the reduction of learning to the varying cue as a result of the strength that accrues to the constant context cue. More generally, the greater the strength of the context cue, the more it reduces the strength of the varying cue.

A second effect of cue interaction explained by R-W concerns the phenomenon of *conditioned inhibition*. It has been shown that a novel cue, I, acquires inhibitory associative strength when it—in combination with a predictive cue, P—is paired with the absence of the outcome. In comparison, a novel cue that by itself is paired with the absence of the outcome acquires zero strength. According to R-W, the combination of P and I is initially expected to produce the outcome (due to the summing of the positive strength of P and the zero strength of I). A discrepancy between the predicted and actual outcomes therefore arises when the combination is paired with the absence of the outcome. This discrepancy leads to a reduction in the strength of I, which therefore becomes negative. (Although the strength of P will also be reduced on such trials, P will regain its strength on other trials in which the outcome continues to be predicted by the occurrence of P in the absence of I.) At asymptote the negative strength of I offsets the positive strength of P, leading to a net expectation of 0 on trials on which the combination of P and I is presented.

Limitations of the R-W Model

Despite its notable successes, the R-W model has several well-known limitations (see Gallistel 1990; Holland et al. 1986; Miller and Matzel 1988). First, whereas the model predicts that the strength of a conditioned

inhibitor should be revised upward toward zero when it is presented alone without reinforcement, in fact such a procedure fails to extinguish the conditioned inhibitor (Zimmer-Hart and Rescorla 1974). Second, the R-W model is unable to account for apparent changes in the associative strength of a cue that occur over a period in which that cue has not been presented (because the model updates only the strengths of cues that are present on a trial). For example, even though presentation of an inhibitor without reinforcement does not reduce its inhibitory power, extinction of the excitatory cue with which the inhibitor had been paired during acquisition can effectively weaken the inhibitor (Kaplan and Hearst 1985; Kasprow, Schachtman, and Miller 1987; Miller and Schachtman 1985). It is as if, when the animal learns that the excitor no longer signals danger, it also loses confidence that the previously paired inhibitor signals safety, even though the inhibitor has not been presented during the extinction phase. Similarly, a cue that was initially "overshadowed" by a more potent excitor will later gain excitatory potential during its absence if the overshadowing cue is extinguished (Kaufman and Bolles 1981; Matzel, Schachtman, and Miller 1985; see also Miller and Matzel 1987). These "indirect" effects on conditioning cannot be explained by R-W, according to which the associative strength of a cue that is absent should not be updated (see equation 3).

Third, the R-W model does not explain the learned irrelevance of a cue that has been randomly paired with an unconditioned stimulus (i.e., the effect) before the effect is made contingent on the cue. Contradicting R-W, the conditioning of such a cue relative to a novel cue is severely retarded. These types of cues are predicted by the model to be equivalent, because they both should begin the conditioning phase with zero associative strength. Fourth, the R-W model predicts that the learning of the novel cue in the blocking paradigm described above will be completely blocked at asymptote. Available empirical results regarding causal induction by humans show, however, that blocking is only partial (e.g., Chapman and Robbins 1990; Shanks 1991; Shanks and Dickinson 1987; Waldmann and Holyoak 1992).

INTERPRETING CONDITIONING PHENOMENA IN TERMS OF CONTINGENCIES

Associative learning models are often contrasted with models based on statistical contingency. Shanks and others (e.g., Chapman and Robbins 1990; but see Chapman 1991) have examined the special case in which the contingency of each of the multiple potentially causal factors that are present is calculated unconditionally over what might be termed the "universal focal set"[2] of all events in the experiment. However, when multiple candidate causal factors are present, contingency for a factor should be computed over subsets of the universal set of events that are

conditional on the constant presence or absence of other factors. We shall now argue that contingency theory, as elaborated with the notion of focal sets in PCM, can account not only for phenomena that have been viewed as major successes for the R-W model, but also for phenomena that contradict the R-W model. Moreover, it provides a framework for understanding when and why the R-W model fails.

Learned Irrelevance

According to PCM, conditioning cue i requires creating a difference between p_i and $p_{\bar{i}}$. The rate at which this difference is created by any single event will be slower for a cue that has been randomly paired with an outcome than for a novel cue, because the cue that had been randomly paired, unlike the novel cue, begins conditioning with large denominators in the two proportions. The impact of any single event in the conditioning phase, therefore, is smaller for the old cue than for the novel cue. Suppose that a cue was present on 100 trials and absent on another 100 trials, and the outcome occurred on 40 trials of each type. The resulting contrast would be zero. Now suppose that, in a subsequent conditioning phase, the outcome always occurs on 5 trials in which this preexposed cue is present, and it never occurs on 5 trials on which this cue is absent. These 10 trials, together with the 200 initial trials, will lead to a contrast of 5/105 (i.e., 45/105 — 40/105). In comparison, for a novel cue, the same 10 trials in the conditioning phase will lead to a contrast of 1 (i.e., 5/5 — 0/5). The result, then, will be a marked attenuation of the rate of conditioning for the preexposed cue relative to a novel cue.

Conditional Contingencies and the Interpretation of Cue Interaction

When multiple potential causal factors are present, philosophers and computer scientists have proposed that assessment of causal relations should not be based on contingencies computed over the universal set of events (Cartwright 1983, 1989; Pearl 1988; Reichenbach 1956; Salmon 1980, 1984; Simpson 1951; Suppes 1970, 1984), because in these situations unconditional contingencies do not reflect what people intuitively judge to be normative causal inferences. In particular, people distinguish between a genuine cause and a spurious cause—a factor that is contingently related to the effect, but is not a cause of it. Unconditional contingencies do not reflect this difference.

These theorists have proposed that, normatively, if a factor is known to be a cause of an effect, then determining the causal status of another factor requires that the contingency of the latter be calculated separately conditional on the presence and on the absence of that cause (a test of "conditional independence").[3] Testing for conditional independence is analogous to comparing experimental conditions to control conditions

in standard experimental design, where extraneous variables are kept constant across conditions. Although this criterion has not been uncontested among philosophers (e.g., Cartwright 1989; Salmon 1984), the prevalent adoption of the analogous principle of experimental design gives an indication of its normative appeal. One important difference between conditional contrasts and comparisons involving experimental design is that conditional contrasts includes observational situations, which generally provide less firm support for causal inferences. In terms of PCM, the adoption of the criterion of conditional contrasts involves computing contrasts for a potential causal factor separately for focal sets that are restricted to events of which the known cause is (a) present, and (b) absent rather than computing them over the universal set of events.

We shall next consider the interpretation of tests of conditional independence, describe a process model for assessing conditional independence, and illustrate the explanation of cue interaction effects according to conditional contingencies in terms of this process model. Let us first consider the interpretation of some possible outcomes of the test of conditional independence for a target factor that has a positive unconditional contingency with the effect (i.e., a possible excitatory cause). For example, suppose we are assessing possible causes of cancer and that smoking cigarettes is an established cause. Now we observe that coffee drinking is also statistically relevant to cancer in that the probability of cancer is higher for people who drink over five cups per day than for those who drink less coffee. However, let us further suppose that people who drink large quantities of coffee also tend to smoke. To tease the influence of coffee drinking apart from that of smoking, it is desirable to calculate the conditional contingency between coffee drinking and cancer separately for cases involving the presence vs. the absence of smoking. The following are four possible outcomes that will be relevant in interpreting blocking and similar cue interaction effects:

Case 1: If both conditional contingencies for the target factor are positive, then the target factor will be interpreted as a genuine cause. For example, if coffee drinking increases the risk of cancer both for smokers and for nonsmokers, then coffee drinking will be interpreted as a genuine cause (unless it turned out to be confounded with some other cause of cancer, such as eating fatty foods).

Case 2: If contingencies for the target factor conditional on both the presence and the absence of the established cause are zero, then that factor will be interpreted as a spurious cause. It is said to be "screened off" (i.e., normatively blocked) from the effect by the conditionalizing cause. For our example, the statistical link between coffee drinking and cancer would be attributed entirely to the confounding between coffee drinking and smoking.

Case 3: If the effect always occurs in the presence of the established cause, regardless of whether the target factor occurs (therefore, the contingency conditional on the presence of the established cause is zero), but the contingency conditional on the absence of the causal factor is positive, then the target factor will be interpreted as a genuine cause. This situation would arise if smoking always caused cancer, so that coffee drinking did not increase the risk of cancer for smokers, but did increase the risk for nonsmokers. In this situation coffee drinking would be interpreted as a genuine cause of cancer. As noted earlier, the zero contingency for a candidate factor (coffee drinking) in the presence of an alternative factor that always produces the effect (smoking) does not give an interpretable estimate of the causal power of the candidate factor. In other words, it would likely be attributed to a ceiling effect (i.e., smoking by itself generates the maximal cancer risk, so that the detrimental impact of coffee drinking is masked for smokers).

Case 4: If the contingency of the target factor conditional on the presence of the established cause is positive, but the effect never occurs in the absence of the established cause (therefore, the contingency conditional on the absence of the established cause is zero), then the two factors will be interpreted as interacting to produce the effect (see equation 2). Such an interaction would exist if coffee drinking in combination with smoking increased the risk of cancer for smokers, but had no effect on the probability of cancer for nonsmokers.

One problem that complicates the test of conditional independence is that the information required for computing the two conditional contingencies is not always available. Recall that in the blocking paradigm a novel cue R is paired with the outcome only when a predictive cue P is also present. Table 12.1 gives a schematic representation of the typical probability of the outcomes for the two cues. The $\overline{P}$ and R cell receives no information, and the outcome always occurs when P is present. Because P is known to have a positive contingency with respect to the outcome, the status of R should be based on conditional contrasts. When the focal set is restricted to events in which P is present (the top row), R has a zero contrast. When the focal set is restricted to events in which P is absent (the bottom row), however, the contrast for R cannot be computed. Because this cue is never presented in the absence of P in this paradigm, $p_{R|\overline{P}}$ is undefined due to division by zero. As Waldmann and Holyoak (1992) noted, because the level of the effect produced by P is already at ceiling, it is

Table 12.1 Probability of the outcomes for cues P and R in the blocking paradigm

	R	$\overline{R}$
P	+	+
$\overline{P}$		0

+ = a positive probability of the outcome
0 = zero probability of the outcome

impossible to determine whether the redundant cue R is a spurious cause (case 2 above) or a genuine cause (case 3). Given that relevant information is missing from the blocking design, subjects who adopt the criterion of conditional independence will be uncertain about the predictive status of the redundant cue, as opposed to being certain that this cue is not predictive, as implied by the R-W learning rule.[4]

It is important to note that there is an asymmetry between the informativeness of tests conditional on the absence vs. the presence of other causes: the tests most likely to clearly rule out a target factor as an independent excitatory cause are those based on the absence of conditionalizing cues. In particular, finding a zero contingency conditional on the absence of other causes clearly rules out a factor as an independent excitatory cause (i.e., it is either spurious, as in case 2, or a component of an interaction contrast, as in case 4), whereas finding a zero contingency conditional on the presence of a known cause is inconclusive (the target might be spurious, as in case 2, but it might instead be genuine, as in case 3). (This interpretation excludes consideration of inhibitory causes, to which we shall return.) Similarly, finding a positive contingency conditional on the absence of other causes constitutes evidence that the cue is an independent excitatory cause (for which case 1 or case 3 might obtain), but a positive contingency conditional on the presence of a known cause could indicate either a genuine independent excitatory cause (as in case 1) or a component of an interactive excitatory cause (as in case 4). Moreover, as noted earlier, the main effect contrast for a candidate factor conditional on the absence of alternative factors gives a better estimate of the causal power of that factor than its contrast conditional on the presence of alternative factors. The fact that tests conditional on the absence rather than the presence of other causes are more informative is reflected in experimental design: if only one type of conditionalizing test can be performed, scientists generally favor designs in which a target factor is manipulated while ensuring that other known causes are absent rather than present. We therefore assume that people will prefer to conditionalize each target factor on the simultaneous absence of all established or likely causes, because this is the test that will be maximally informative.

The above analyses of the informativeness of conditional contingency tests apply in the case of possible excitatory causes, but not in that of possible inhibitory causes. A test of a target factor in the absence of all established causes cannot demonstrate that the factor is an inhibitor because, unless some excitatory cause is operating, the impact of an inhibitor will be obscured by a cellar effect. That is, if the outcome is not being produced by some excitatory cause, an inhibitor cannot achieve a nonzero contingency. We assume that, as a general principle based on a preference for cognitive simplicity, a factor will not be deemed causal unless positive evidence of a causal interpretation is obtained. Accordingly, the default interpretation of a zero contingency is that the factor is noncausal (rather than inhibitory). This assumption is supported by the fact that simply presenting a cue alone without reinforcement, while another cue presented alone is reinforced, generally does not yield strong conditioned inhibition (Baker 1977). The former cue has a negative unconditional contingency, but its contingency conditional on the presence of the latter cue cannot be computed due to the lack of information on the frequency of the effect when both cues are present. Thus for a candidate inhibitory factor the most informative tests will involve computation of its contingency conditional on the presence of a single excitor, coupled with the absence of all other known causal factors. If there is more than one known excitor, it will be desirable to perform separate tests for the candidate factor conditional on the presence of each excitor in turn. If the candidate yields a negative contingency conditional on the presence of an excitor, it will be interpreted either as a main effect inhibitory cause or as a component of an inhibitory interaction.

Conditioned Inhibition and "Indirect" Extinction of Associative Strength

PCM can account both for the acquisition of conditioned inhibition and for the failure to extinguish a conditioned inhibitor by presenting it alone without the outcome. Table 12.2 schematically represents the typical probability of the outcomes for the two cues in the learning phase of the conditioned inhibition paradigm. When P is presented alone, the outcome occurs, but when P and I are presented in combination, the outcome does not occur. No information is received about the $\bar{P}$ and I cell (the empty one in the table) during the learning phases. Notice that the set of events so far shows a negative conditional contrast for I conditional on the presence of P (i.e., $p_{I|P} < p_{\bar{I}|P}$). Therefore, PCM predicts that I will become inhibitory. Now consider an extinction phase in which I is presented alone without the outcome. The $\bar{P}$ and I cell will be filled in with the information that the probability of the outcome in the presence of I alone is zero. This information will have no impact on the crucial conditional contingency—that of I in the presence of P—and hence will not yield extinction (Zimmer-Hart and Rescorla 1972).

Table 12.2 Probability of the outcomes in the learning phase of the conditioned inhibition paradigm

	I	$\bar{I}$
P	0	+
$\bar{P}$		0

+ = a positive probability of the outcome
0 = zero probability of the outcome

In addition to correctly predicting that inhibition cannot be extinguished directly by presenting the inhibitory cue without reinforcement, PCM can account for results demonstrating that conditioned inhibition can be extinguished indirectly, by extinguishing the excitatory strength of the cue with which the inhibitor had been paired (Kaplan and Hearst 1985; Kasprow, Schachtman, and Miller 1987; Miller and Schachtman 1985). As we explained, this counterintuitive finding contradicts the R-W model. In this indirect extinction procedure, P rather than I is presented alone without the outcome. Reducing the frequency of the outcome in the P and $\bar{I}$ cell reduces the magnitude of the negative contrast for I conditional on the presence of P, and hence diminishes the perceived inhibitory impact of I (e.g., Miller and Schachtman 1985).

A Process Model for Assessing Conditional Dependence and Independence

We have so far been describing contingency theory at the computational level of analysis. Here we will describe an algorithmic instantiation of the theory. This process model is based on PCM, with extensions to specify which conditional contingencies are computed. Contingency analysis can of course be evaluated independently of this particular instantiation, but this model will serve to provide a detailed illustration of how contingency theory might account for cue interactions.

A plausible psychological model of causal inference based on contingency analysis must specify mechanisms that would allow people to decide (a) what cues should be used to conditionalize others, (b) what conditional tests to perform once a set of conditionalizing cues has been selected, and (c) how to integrate the resulting contingency information to make causal assessments of the cues. In situations in which there is no guidance from prior knowledge, every cue is potentially causal. Given n binary cues, exhaustively conditionalizing the contingencies for each target cue on every combination of the presence and absence of the other cues requires computing $2^{n-1} \cdot n$ contingencies. Given processing limitations, it is crucial to specify how people select which contingencies to compute. It is also likely that many of the cue combinations that would be relevant to a contingency analysis will never actually occur. Accordingly, it is necessary to specify which contingencies will be computed in the face of missing information.

Let us first consider the selection of conditionalizing cues. The ideal set of conditionalizing cues will include all those and only those that are actually causal. Given the limitations of knowledge, the best people can do is to select as conditionalizing cues those they currently believe to be plausible causes. In cases in which prior knowledge is relevant, such knowledge will be used to establish certain cues as likely causes, and the contingencies for other cues will then be conditionalized on the (perhaps tentatively) established causes. If such prior knowledge is lacking, people may nonetheless use some heuristic criterion to select an initial set of conditionalizing cues. A simple heuristic that might be employed is to include any cue that is noticeably associated with the effect. That is, people may follow the tacit rule: If the effect is likely to occur when the cue occurs, tentatively assume that the cue may be causal. Contingencies are not computed in this initial phase of selecting conditionalizing cues; rather, people simply identify a pool of cues that have been paired with the effect, which will be treated as an initial set of plausible causes. There is some evidence of such an initial phase of cue selection based on positive associations. For example, Rescorla (1972) found that a cue that was randomly paired with the outcome (i.e., one that was associated with the outcome but noncontingent with it) appeared to initially acquire associative strength, which eventually disappeared after several sessions of training. The association heuristic suggested here implies that this phase implicitly ignores the possibility of cues' being interactive or inhibitory causes. The sole presence of an inhibitory cause, for example, will be perceived as a lack of association.

Contingency assessment will occur in the subsequent phase, in which people will compute the conditional contingencies of all cues based on the set of conditionalizing cues identified in the initial phase. In Cheng and Novick's (1990) terminology, the set of conditionalizing cues defines the focal sets for contingency computations. The initial set of conditionalizing cues can be dynamically updated if contingency assessments indicate that cues that at first appeared to be plausible causes are in fact spurious or that cues initially viewed as causally irrelevant are in fact causal. That is, after conditional contrasts are calculated based on the initial set of conditionalizing cues, these contrasts will be used to update that set of cues. Cues in the set that have zero or low contrasts may be dropped, and other cues outside the set that have noticeable positive or negative contrasts may be added. Changes in the set of conditionalizing cues will in turn change the relevant conditional contingencies for all cues, which may alter subsequent causal assessments. The entire assessment process will thus be iterative. If the values of the cues stabilize as the process iterates, the process will return these values and stop. Otherwise, the process will stop after an externally determined number of iterations.

In assessing conditional contingencies, heuristics will be required to determine which tests (of those possible given the cue combinations that are actually presented) should in fact be performed. We assume, based on the arguments presented earlier, that people will prefer to conditionalize the contingency for each target factor on the simultaneous absence of all conditionalizing cues. If this is not possible, then they will try to select a focal set in which as many conditionalizing cues are absent as possible, while the rest of the conditionalizing cues are constantly present.

In general, application of the contingency analysis will necessarily be constrained by the information actually provided by observation.

In addition to specifying what cues are selected to form the conditionalizing set and which conditional contingencies are computed, a process model must specify a response mechanism that translates the calculated contingencies into causal judgments. If all conditionalizing cues can be kept either absent or present and there are no ceiling effects for excitatory cues, the confidence associated with the contrast values based on these focal sets will be relatively high. But if the experimental design omits cases that would be relevant in assessing the conditional dependence or independence of a target factor such that there are ceiling effects or some of the conditionalizing cues cannot be kept constant, the confidence associated with the contrast values based on these focal sets will be relatively low. In such experiments, if subjects are not given the choice of withholding judgment, they may base their causal assessments on a mixture of the best available focal sets—for example, the unconditional as well as the conditional contingencies for cues. Mean ratings over subjects may therefore reflect some mixture of the evidence provided by conditional and unconditional contingencies.

When subjects do not all use one and the same focal set to compute contingencies, the mean causal judgment about a cue (averaged across subjects in an experimental condition) should reflect some mixture of assessments based on the multiple focal sets used. These may include the universal focal set of all events in the experiment (i.e., unconditional contingencies) and various more restricted focal sets (i.e., conditional contingencies). The response mechanism must then account for how multiple contingencies are integrated. The clearest situation is that in which the relevant unconditional and conditional contingencies for a factor are all computable and equal to zero, in which case subjects should be certain that the factor is noncausal. Beyond this limiting case, we make no claim about the exact quantitative mapping between contingency values and subjects' responses. Our assumption is that subjects' causal estimates will increase monotonically with a nonnegatively weighted function of the contingency values of their focal sets. Individual subjects may compute and integrate multiple contingencies for a cue (e.g., by simple averaging). Alternatively, each subject may use only one focal set, but different subjects may use different focal sets, in which case the mean ratings may mask distributions that are in fact multimodal. We will refer to the assumption that causal ratings may be based on multiple contingencies (calculated either by individual subjects or by different subjects) as the "mixture-of-focal-sets" hypothesis. As we will see, this hypothesis helps to explain circumstances in which partial rather than complete blocking is observed.

Computing contingency conditional on the presence of an alternative cause raises the problem of how an alternative cause that happens to be

constantly present in the current focal set can be distinguished from an enabling condition. To distinguish between them, Cheng and Novick (1992) refined their definition of an enabling condition as follows. Let i be a factor that is constantly present in the current focal set. Factor i is an enabling condition for a cause j in that focal set if i covaries with the effect in another focal set and j no longer covaries with the effect in a focal set in which i is constantly absent. In contrast, i is an alternative to cause j if i covaries with the effect in another focal set and there exists a focal set in which i is constantly absent, but j continues to covary with the effect in that set.

To summarize, our proposed process model assumes that subjects will (a) identify as initial conditionalizing cues those that are noticeably associated with the effect; (b) compute contingencies for each target factor conditional on the absence of as many conditionalizing cues as possible, dynamically revising the set of conditionalizing cues in the process; and then (c) use the computed conditional contingencies and/or unconditional contingencies to produce causal assessments for the cues.

Interpreting Blocking, Partial Blocking, and Other Cue Interaction Effects

We will now consider how generalized contingency theory, in particular as implemented in the process model we have described, can account for blocking, partial blocking, overshadowing, retroactive extinction of overshadowing, and other cue-interaction effects.

Blocking and Partial Blocking

In the standard blocking design illustrated in table 12.1, the unconditional contingency is higher for the predictive cue P than for the redundant cue R (because the outcome sometimes occurs in the absence of R, but never in the absence of P), although the contingency is positive for both. Thus even subjects who compute contingency over the universal focal set would be expected to show at least partial blocking (i.e., the higher response strength for P than R, both strengths being positive). It is possible, however, to design an experiment in which unconditional contingency is held constant for two cues, and yet their causal statuses differ. Such designs have been used in classical conditioning experiments, as well as in experiments on causality judgments by humans (Chapman and Robbins 1990, experiment 1; Shanks 1991, experiment 2). The design used by Shanks is schematized in table 12.3. After being presented with a series of "case histories" (patterns of patients' symptoms associated with various fictitious diseases), subjects were asked to rate how strongly they associated each symptom with each disease using a 0–100 rating scale. In what Shanks termed the "contingent" set, the compound cue AB signaled the presence of disease 1 (15 trials), but symptom C by itself did so as well

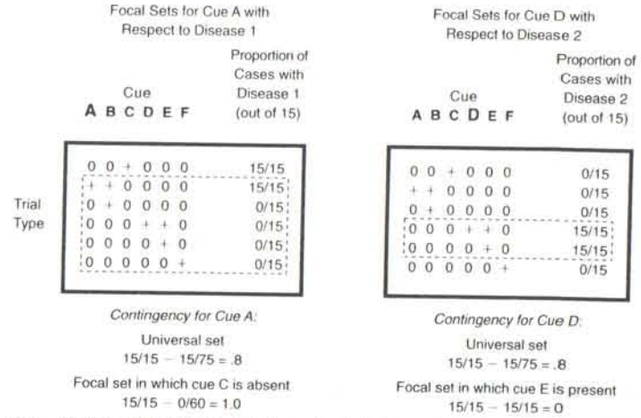
Table 12.3 Conditions, trial types, number of trials, and percentage of correct diagnoses for experiment 2 of Shanks (1991). (Adapted from Shanks 1991.)

Condition	Trial type	Trials	% correct
"Contingent"	C → D1	15	100
	AB → D1	15	100
	B → 0	15	94
"Noncontingent"	DE → D2	15	100
	E → D2	15	100
	F → 0	15	94

(15 trials). However, cue B by itself signaled the absence of the disease (15 trials), as did the absence of A, B, and C (45 trials). In the "noncontingent"[5] set, compound cue DE signaled the presence of disease 2 (15 trials), as did the presence of cue E alone (15 trials). In contrast, cue F alone signaled the absence of the disease (15 trials), as did the joint absence of D, E, and F (45 trials).

The critical comparison is between the association rating given to symptom A for disease 1 and the association rating given to symptom D for disease 2. Although the contingency computed over the entire set of events presented for both relations is .8 (see figure 12.1), the R-W model predicts that, because D is paired with a better predictor, E, subjects should rate D as less associated than the corresponding symptom A, which is paired only with a nonpredictor, B. This difference was observed. In other words, the rating given to a cue was reduced if a competing cue was a better predictor of the relevant disease, even through unconditional contingency was equated. But although subjects gave higher mean ratings to A than to D (59 vs. 34, respectively), even cue D received modestly positive ratings, whereas the R-W model predicts that at asymptote the strength of the association between D and disease 2 should be 0 (see Melz et al. 1993). Shanks's experiment is representative of several other cases in which human subjects show only partial blocking rather than complete blocking as the R-W model would predict (e.g., Chapman and Robbins 1990; Shanks and Dickinson 1987; Waldmann and Holyoak 1992).

A contingency analysis of this and another cue competition experiment by Shanks is provided by Melz and co-workers (1993). Figure 12.1 illustrates the computation of contingencies for the cues crucial for comparison, A in the "contingent" set and D in the "noncontingent" set. As shown in the figure, the unconditional contingency (i.e., the contingency computed over the universal set of all events) is .8 for critical cue A with respect to disease 1, as is that for cue D with respect to disease 2. To test conditional independence of these cues with respect to the particular disease, we apply the process model described above. With respect to disease 1 (see the left half of figure 12.1), only cues A, B, and C will be identified as initial conditionalizing cues, because these are the only cues that are ever accompanied by disease 1. Cue B has a contingency of 0 in



Note. Letters A to F denote cues. Solid-line rectangles indicate universal focal sets; dashed-line rectangles indicate conditional focal sets. Large bold letters denote the crucial cues for comparison.

Figure 12.1 Potential focal sets in Shanks's (1991) experiment 2. (From Melz et al. 1993.) *Note:* Letters A to F denote cues. Solid-line rectangles indicate universal focal sets; dashed-line rectangles indicate conditional focal sets. Large bold letters denote the crucial cues for comparison.

the focal set, from which both A and C are absent (rows 3–6). Cue C has a conditional contingency of 1.0 in the focal set, from which cues A and B are both absent (rows 1 and 4–6). Each of the remaining cues (D, E, and F) has a conditional contingency of 0 in the focal set, from which all conditionalizing cues (A, B and C) are absent (rows 4–6).

The contingency for cue A conditional on the absence of both B and C cannot be computed, because A does not occur in the absence of B. However, A has a contingency of 1.0 in the focal set, in which B is present and C is absent (rows 2–3). From the first iteration of conditional contingency assessment, it follows that B will be assessed as noncausal and be dropped from the set of conditionalizing cues, so that only A and C will remain as conditionalizing cues. The relevant contingency for A then becomes that which is conditional on the absence of C (rows 2–6, enclosed by a dashed rectangle in the illustration) and has a value of 1.0. This is equal to the value of the relevant conditional contingency obtained for A in the previous iteration. As is the case for A, none of the values of the relevant conditional contingencies for any of the other cues change as a result of dropping B from the conditionalizing set.

For disease 2 (see the right half of figure 12.1), cues D and E will be selected as conditionalizing cues. Because D never occurs in the absence

of E, its contingency can be calculated conditional only on the presence of E. For this focal set (enclosed by the dashed rectangle in the illustration), the conditional contingency for D with respect to disease 2 is 0. The difference between the computed contingency for cue A with respect to disease 1 (1.0) and that for cue D with respect to disease 2 (0) provides an explanation for cue competition—the lower ratings given to D than to A. In addition, cue E has a contingency of 1.0 conditional on the absence of D (rows 1–3 and 5–6 in the right half of figure 12.1). All other cues have a contingency of 0 with respect to disease 2 in the absence of cues D and E.

Now consider how partial blocking might arise. As we mentioned, the R-W model predicts that associative learning of a novel cue in the blocking paradigm will be completely blocked at asymptote; yet all available empirical results regarding humans show that blocking is not complete. The above contingency of 0 for cue D was conditional on the presence of cue E. However, in the presence of E the effect always occurs. Since it was not possible to conditionalize the contingency for D on the absence of E, subjects should be uncertain of the interpretation of the contingency value of 0. Accordingly, at least some subjects may assess the unconditional contingency for D (i.e., over the universal set of events), which is 0.8. Assuming that subjects' causal ratings reflect a mixture (either within individual subjects or across subjects) of these two contingencies, D will receive a relatively low but positive mean rating. That is, it will be partially blocked. Moreover, the prediction of cue competition remains, since the contingency for A (1.0) is still higher than the mixture of the contingencies for D (0.8 and 0). In sum, for situations in which there is no focal set that allows an unambiguous interpretation, if there is a mixture of focal sets either within subjects or across subjects, contingency theory predicts partial blocking in addition to cue competition.

Overshadowing and Retroactive Reduction of Overshadowing

A similar contingency interpretation can be provided for experiments that have demonstrated that a salient predictive cue acquires greater strength than a less salient cue that is perfectly correlated with it (i.e., the salient cue overshadows the less salient cue) and that extinguishing the salient competing predictor can increase the excitatory power of the previously overshadowed cue (Kaufman and Bolles 1981; Matzel, Schachtman, and Miller 1985).[6]

When two cues are perfectly correlated with each other, the association of the salient cue with the outcome is likely to be noticed earlier than the association of the less salient cue. Accordingly, the former cue will be selected earlier than the latter as a conditionalizing cue. It follows that the subject will initially attempt to conditionalize the contingency of the nonsalient or "pallid" cue on the state of the salient cue, but not vice versa. But due to the absence of information in this design regarding the occurrence of the outcome in the presence of one cue and the absence of the other cue, neither of the relevant contingencies for the pallid cue (i.e., those conditional on the presence and on the absence of the salient cue) can be computed. Accordingly, the subject will be uncertain about the causal status of the pallid cue during this phase. Meanwhile, given the positive unconditional contingency of the salient cue with respect to the outcome, this cue will be judged causal. Hence, it will be confirmed as a conditionalizing cue for computing the contingency of the pallid cue, whereas the pallid cue may never acquire that status with respect to the salient cue. In the phase that ensues in a retroactive paradigm, however, the salient cue is presented alone (and it is not followed by the outcome). Information then becomes available for computing the contingency of the pallid cue conditional on the presence of the salient cue. The resulting positive value of this conditional contingency predicts the increased causal strength of the pallid cue.[7]

Direction of Causality

The above analyses of Shanks's results assume that subjects based their inferences on calculations that were in turn based on probabilities of diseases conditional on the various symptoms. This seems likely for at least some subjects in view of the instructions and the learning procedure. The instructions did not make it clear, for example, whether a disease name referred to the cause of the associated symptoms or was simply a label for them. However, if the causal direction is made salient to subjects, then the predictions of the R-W versus contingency approaches are very different indeed. The R-W model, although often interpreted as an account of causal induction, does not in fact draw any distinction between a context in which cues are interpreted as possible causes of an effect (the typical situation involving *predictive* learning) and a context in which cues are interpreted as possible effects of a common cause (*diagnostic* learning). Diagnostic tasks require reasoning in a backward causal direction (e.g., from symptoms, which are effects, to underlying diseases, which are interpreted as causes of the symptoms).

Waldmann and Holyoak (1992) have shown that the degree of cue competition is radically different depending on whether people interpret the cues as the causes of an effect to be predicted or as the effects of a cause to be diagnosed. In their experiment 3, Waldmann and Holyoak exposed subjects to a series of trials in which states of previously unfamiliar cues (buttons connected to an alarm system) were paired with states of the alarm system. Each button had two settings, on and off, as did the alarm system. Subjects in a predictive condition were told that pressing one or more buttons would cause the alarm to go on. In this condition the states of the buttons were thus characterized as possible causes, and the states of the alarm system were characterized as possible effects. In contrast, subjects in a diagnostic condition were told that one or more of

the buttons signaled whether the alarm system was on. Notice that the direction of causality was reversed in this cover story relative to that in the cover story for the predictive condition. As in the predictive condition, however, subjects saw only the state of the buttons. They had to respond by predicting the state of the alarm, and then they received feedback as to the actual state of the alarm. The cues presented and the responses required were thus equated across the conditions.

The experimental design in both conditions included two phases corresponding to a standard blocking paradigm. Phase 1 established a certain button (P) as a perfect predictor of the state of the alarm. A second button (C) was constantly set to the value off, and a third button (U) varied in a fashion that was uncorrelated with the state of the alarm. Phase 2 maintained these same contingencies, but also added a fourth button (R) that was always on when P was on and off when P was off. Thus, if subjects learned to predict the state of the alarm from the states of the buttons according to the R-W rule, in both conditions learning should have have been blocked in phase 2 for button R by the associative strength that would already have accrued to button P in phase 1.

After each phase of the design, subjects in both conditions rated the degree to which the state of each button was predictive of the state of the alarm using a scale from 0 to 10, where 10 indicated that the cue was a perfect predictor and 0 indicated that the cue was not a predictor. As would be expected on the basis of both contingency theory and the R-W model, the ratings obtained after phase 1 (panel A of figure 12.2) indicated that in both the predictive and diagnostic conditions button P was established as a strong predictor of the state of the alarm, whereas both cues C and R were rated as very weak predictors.

The most important findings involve the predictiveness ratings obtained after phase 2 of the experiment (panel B of figure 12.2). According to the R-W model, the associative strength acquired for the redundant button R should have approached 0 in both the predictive and diagnostic conditions (as should also have happened for the noncontingent buttons C and U). That is, associative learning for cue R should have been entirely blocked by the prior strength of cue P. However, a very different prediction follows from causal contingencies. If people tend to compute contingencies from causes to effects rather than from effects to causes—even when the causal direction is opposite to the order of cue-outcome presentation—then contingency theory predicts that no blocking will be observed in the diagnostic condition. In the diagnostic context the redundant cue, button R, is not an alternative possible cause, the contingency of which should be conditionalized on the status of the established predictor, button P; rather, the state of R is simply a second possible effect of the same cause. If alternative effects, unlike alternative causes, are given separate contingency analyses, then no cue competition should be observed. And indeed, Waldmann and Holyoak found that, while button P was rated

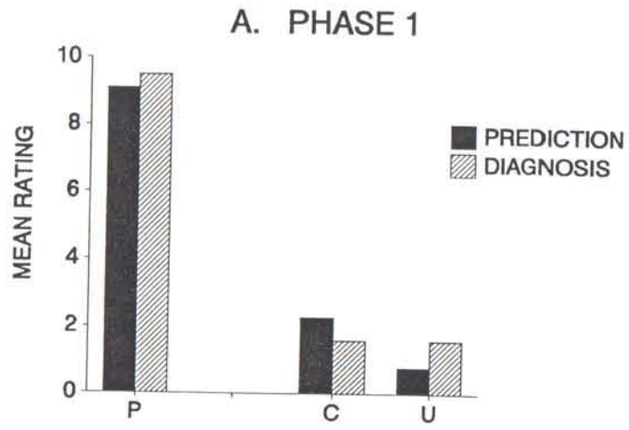


Figure 12.2 Mean predictiveness ratings for predictive and diagnostic conditions obtained in phase 1 (panel A) and phase 2 (panel B) of Waldmann and Holyoak's experiment 3 for the initial predictive cue (P), the redundant predictive cue (R), the constant uncorrelated cue (C), and the varying uncorrelated cue (U). (From Waldmann and Holyoak 1992.)

higher than button R in the predictive condition (9.7 and 4.3, respectively), in the diagnostic condition buttons P and R were given high and statistically equal ratings (9.6 and 8.7, respectively). This interaction between causal direction and the difference in the mean ratings for buttons P and R was highly significant. In addition, the results indicated that, even in the predictive condition, blocking for button R was only partial: the rating for cue R was significantly higher than the ratings for the noncontingent cues C and U. The latter finding is consistent with other evidence that blocking is only partial in human causal induction, as we discussed earlier.

In sum, evidence from studies of human causal induction using paradigms formally similar to blocking studies in animal conditioning has revealed phenomena that are inconsistent with the predictions of the R-W model, but interpretable in terms of a contingency theory such as PCM.

UNDERSTANDING FAILURES OF THE R-W MODEL FROM A CONTINGENCY PERSPECTIVE

Our analysis indicates that contingency theory, as generalized by PCM to apply to various focal sets, including ones for computing conditional contrasts, can account for a wide range of phenomena that are obtained in studies of animal conditioning and human causal induction—not only those that are explained by the R-W model, but also those that are problematic for that model. In addition, our process model makes a number of predictions that have yet to be tested. For example, this model predicts when the causal status of a cue can be uncertain and when the distribution of judgments regarding a cue can be multimodal.

One of the primary attractions of the R-W model is the apparent simplicity and generality of its learning algorithm. However, the simplicity of the model can be questioned (Gallistel 1990); and whether or not it is simple, its wide range of empirical shortcomings indicates that it is simplistic. It may be instructive to consider when and why the R-W model fails to account for phenomena concerning conditioning and causal induction.

First, the R-W model does not represent cause-absent information—in particular, the proportion of trials on which the outcome occurs in the absence of a cue. To understand why the lack of representation of the cause-absent proportion is a weakness, let us first consider the reason for the model's successes. On the basis of interaction among cues that are defined solely in terms of their presence, the model is able to account for a number of apparent effects of cause-absent information: it accounts for the role of contingency (Rescorla 1968), the acquisition of conditioned inhibition (e.g., Chapman and Robbins 1990), blocking (e.g., Rescorla 1981), and other cue interaction effects (e.g., Wagner et al. 1968). In each of these cases, two or more cues that had an identical cause-present proportion, but a different conditional or unconditional cause-absent proportion, have been observed to elicit different behavior, as predicted by the model.

Two properties of the model allow it to arrive at these predictions. First, the model indirectly tallies the cause-absent proportion with respect to a target cue in terms of the cause-present proportion of one or more other cues that are present when the target cue is absent; these surrogate cues therefore acquire weights that reflect the cause-absent proportion of the target cue. In some applications of the model, the surrogate cue is one that represents the context, which is constantly present (and hence is present on occasions when the target cue is absent). Second, on trials when the two cues are both present, the strength of the target cue is adjusted toward the difference between the cause-present proportion of the target cue and the cause-present proportion of the surrogate cues (i.e., the cause-absent proportion of the target cue), potentially yielding contingency as the asymptotic output. In sum, R-W relies on the pairing of cues to transmit the indirectly tallied cause-absent proportion.

Cheng and Novick (1995) present a derivation of when the R-W model does and does not compute conditional contingencies at asymptote. Their analysis shows that it does so for a type of design with multiple cues in which every combination of cues except the one with a single cue can be characterized as a proper superset of all sets with fewer cues (i.e., the cue combinations are nested). In such designs, the strengths of the cues in each combination sum to the relative frequency of the outcome for that combination, implying that, for any combination with multiple cues, the strength of the cue in it that does not belong to the next smaller combination is equal to the contingency of that cue conditional on the presence of the cues in the smaller combination (i.e., the rest of the cues in the larger combination).

Cheng and Novick (1995) also present a derivation of the conditions under which conditional contingencies estimate the causal power of a cue. Their analysis of the R-W model and of conditional contingencies shows that in some nested designs the conditional contingencies computed by the R-W model give an estimate of causal power, whereas in others the conditional contingencies computed by this model do not give such an estimate. Those situations in which the R-W model estimates causal power include those represented by Kamin's (1969) blocking design, unconditional contingency (Rescorla 1968), and the acquisition of conditioned inhibition. In these situations, the R-W model is successful in predicting the observed results (see Cheng and Novick 1995, however, for an explanation of the partial success of the R-W model in predicting the amount of blocking). Those designs in which the R-W model does not estimate causal power include the extinction of conditioned inhibition, retroactive unblocking, and the retroactive reduction of overshadowing (see Miller and Matzel 1988). In these situations, the R-W model fails to predict the observed results.

A second problem with the R-W model is that the causal or conditioning strength of a cue with respect to an effect is represented by a single

parameter—the associative strength of the link between the cue and the outcome. The model therefore loses information about sample size, leading to its failure to account for learned irrelevance and, more generally, people's sensitivity to reliability as a function of sample size (Koslowski et al. 1989; Nisbett et al. 1983). Moreover, the R-W model does not offer any way to represent the difference between lack of certainty about a causal association and high certainty that such an association has some medium strength. In contrast, the outcome of a contingency analysis can include not only a definite evaluation of the causal status of a cue, but also uncertainty about its status. Uncertainty naturally falls out of PCM when a relevant contingency is not computable, as in the case of the redundant cue in the blocking paradigm.

For the same reason, the R-W model cannot account for causal assessments that result from comparing the distinct causal status that a cue has in different focal sets. In particular, the R-W model cannot represent the distinction between a cause and an enabling condition or that between an enabling condition and a causally irrelevant cue. This deficit arises because the status of an enabling condition results from the cue's being causal in one focal set and having a noncomputable contingency in another focal set.

This last point brings up the related problem of the need to specify (potentially multiple) focal sets. Our explanations of enabling conditions and of partial blocking provide examples of the use of such an assumption. One might ask, Will the R-W model be able to explain these phenomena if it is amended with the assumption of computation over multiple focal sets? With respect to blocking (see table 12.1), R-W predicts that a redundant cue, R, should have zero associative strength regardless of which focal set is adopted. For none of the focal sets that arise in a contingency analysis is there ever a discrepancy between the expected outcomes based on R-W and the target outcomes for any trial on which R is present. (See the appendix for derivations of the asymptotic weights of the cues assuming various focal sets.) Considering either the focal set in which the predictive cue is always present (i.e., the top row in Table 12.1) or the universal focal set (i.e., the entire table), the outcome is completely predicted by P. Considering the focal set in which the predictive cue is always absent (i.e., the bottom row), R is never present. Therefore, the strength of R is never revised from zero. In sum, even when amended with the concept of a focal set, the R-W model, unlike our process model, cannot predict the partial blocking of R. Nor can it predict a possible multimodal distribution of judgments regarding R. With respect to the status of an enabling condition, because the R-W model does yield a definite value of strength for a constant cue, it cannot yield the uncertainty that leads to the reliance on the status of the cue in another focal set. Even if the model is applied to a focal set in which the cue is constant and another in which it varies, the result will be two strengths for that cue. It is not clear how this result (i.e., an enabling condition represented by two strengths) can be distinguished from that involving a cause that has different strengths in different focal sets.

Finally, the R-W model does not provide any way to distinguish the case in which cues are possible causes (predictive learning) from that in which cues are possible effects (diagnostic learning). That is, cues and outcomes are defined with respect to their roles as stimuli presented vs. responses made rather than with respect to their conceptual roles as causes or effects. However, a cause can be either a stimulus or a response (as can an effect). As a result, the R-W model is unable to explain interactions between perceived causal direction and cue competition (Waldmann and Holyoak 1992).

It is not obvious to us that any of the above shortcomings of the R-W model can be readily amended. Contingency theory provides a basis for formulating alternative models of how natural adaptive systems operate as intuitive statisticians.

APPENDIX: ASYMPTOTIC WEIGHTS OF A NETWORK WITH A BLOCKING DESIGN OBTAINED BY APPLYING THE R-W MODEL TO VARIOUS FOCAL SETS

We show that amending the R-W model with the same assumptions regarding focal sets as those made by contingency theory does not change its prediction of complete blocking.

Deriving Asymptotic Weights

Because our derivation is based on asymptotic weights, we first describe a method for deriving such weights. To obtain the asymptotic weights of a network according to the R-W model, we note the equivalence between the R-W learning rule and the least-mean-squares (LMS) rule of Widrow and Hoff (1960) (cf. Sutton and Barto 1981). The Rescorla-Wagner/Widrow-Hoff rule implements an iterative algorithm for computing the solution to a set of linear equations defined by the set of stimulus-response patterns presented to the network. A pattern is a configuration of stimuli and responses deterministically describing a set of trials. If the input stimulus patterns are linearly independent, then the updating rule will discover a unique solution. Even if the stimulus patterns are not linearly independent, the network will still converge provided that the learning rate is sufficiently small and that the various stimulus patterns occur with sufficient frequency in the input sequence. The network will converge so as to minimize the sum of the squared errors over the stimulus patterns. That is, the equation

$$\mathbf{E} = \sum_{p} \pi_p t_p \left(\lambda_p - \sum_{i} V_{pi} \right)^2 \qquad (4)$$

will be minimized, where p is the index for a particular stimulus-response pattern, π_p is the frequency of pattern p, ι_p *is the learning rate associated with pattern* p (β_i and γ_i, respectively, for the presence and the absence of outcome j), λ_p is the desired output for the outcome of the pattern (usually either 0 or 1), and

$$\sum_i V_{pij}$$

is the actual output for the pattern, which is equal to the sum of the weights V_i associated with every present cue i for the pattern. If the reinforcement learning rate β_i is equal to the nonreinforcement learning rate γ_i, the ι_p term may be omitted from the equation. We assume that the learning rates β_i and γ_i are equal in the rest of this chapter. Thus the asymptotic weights of a network according to the R-W model can be calculated analytically by minimizing the sum of the squared errors given by equation 4. This minimum value may be obtained by setting the partial derivatives with respect to each weight to 0 and solving the resulting set of equations.

A Predictive Cue and a Redundant Cue

First, consider the network corresponding to the design summarized by table 12.1. There are three patterns of trials: When P and R are both present, the outcome always occurs; when P is present and R is absent, the outcome also always occurs; but when P and R are both absent, the outcome never occurs. Because the π_p and ι_p terms do not affect our result, for simplicity of exposition we assume that the three patterns occur with equal frequency and that the learning rates for the various patterns are equal, so that we omit π_p and ι_p from equation 4. Applying the equation to all events in the table (the universal focal set),

$$\mathbf{E} = [1 - (V_P + V_R)]^2 + (1 - V_P)^2.$$

We see that **E** will have its lowest value when $V_P + V_R = 1$ and $V_P = 1$. Therefore, the asymptotic solution for this network is $V_P = 1$ and $V_R = 0$. That is, the redundant cue R is completely blocked. Note that the pattern involving the joint absence of P and R does not lead to any error terms, because no cue is present. Therefore, applying the R-W model to the focal set consisting of trials in which P is always present yields the same asymptotic solution for V_P and V_R.

Adding a Constant Context Cue

Applications of the R-W model often assume that there is a constantly present context cue, K. The error for the universal set is then

$$\mathbf{E} = [1 - (V_P + V_R + V_K)]^2 + [1 - (V_P + V_K)]^2 + (0 - V_K)^2.$$

We see that **E** will have its lowest value when $V_K = 0$, $V_P = 1$, and $V_R = 0$. That is, R is completely blocked. If we adopt the focal set consisting of trials on which P is constantly present, we drop the third error term above, obtaining

$$\mathbf{E} = [1 - (V_P + V_R - V_K)]^2 + [1 - (V_P + V_K)]^2.$$

By inspection, **E** will be at a minimum when

$$V_P + V_R + V_K = 1 \tag{5}$$

$$V_P + V_K = 1. \tag{6}$$

There is no unique solution of V_P and V_K in this case. But subtracting equation 6 from equation 5 yields the solution $V_R = 0$. Thus R is completely blocked in this as well as in all of the above networks.

NOTES

Preparation of this paper was supported by NSF Grant DBS 9121298 to Cheng and by a UCLA Academic Senate Research Grant to Holyoak. We thank Michael Waldmann and Eric Melz for their extensive contributions to the research reviewed as well as for their helpful discussions of the chapter. Herbert Roitblat provided valuable comments on an earlier draft. Requests for reprints may be sent to Patricia Cheng at the Department of Psychology, Franz Hall, University of California, Los Angeles, California 90095-1563.

1. This definition of a two-way interaction contrast differs from the one proposed by Cheng and Novick (1990), which does not contain the product terms and is therefore less normative.

2. What is referred to here as the "universal set" is actually the pragmatically restricted set of events that occur in the conditioning experiment (i.e., a small subset of the "truly" universal set of all events known to the subject). This contextual delimitation of the largest relevant focal set implies that even the cases in the "cause-and-effect-both-absent" cell are restricted to a small finite number.

3. When there are multiple known causes, assessing the status of a potential causal factor normatively requires computing its contingencies while exhaustively conditionalizing on every combination of the presence and absence of the other cues. We do not mean to imply that a test of conditional independence is the only process for differentiating between genuine and spurious causes (Lien and Cheng 1992).

4. The prediction of uncertainty does not generalize to situations in which the representation of the target phenomenon does not have a maximum value, as does the probability of a phenomenon.

5. Because the critical cues in Shanks's "noncontingent conditions" were contingently related to the respective diseases by the conventional definition, the labels for his stimulus sets in experiments 1 and 2—"contingent condition" and "noncontingent condition"— do not conform to conventional usage.

6. It should be noted, however, that analogous conditioning experiments with animals that attempted to find indirect effects of increasing (rather than decreasing) the excitation of a previously paired cue have failed to obtain such effects (see Miller and Matzel 1988). However, "retroactive blocking"—reduction in the causal value of one cue as a result of increasing the apparent predictiveness of another cue with which it had been previously paired—has been observed in studies of causal induction by humans (Chapman 1991; Shanks 1985). These effects, however, have been relatively small in magnitude.

7. Consideration of the unconditional contingency for the pallid cue yields the same prediction. As the salient cue becomes extinguished, it no longer maintains its conditionalizing status. The positive unconditional contingency of the pallid cue then becomes interpretable as evidence that the latter is in fact causal.

REFERENCES

Alloy, L. B., and 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.

Baker, A. G. (1977). Conditioned inhibition arising from a between-session negative correlation. *Journal of Experimental Psychology: Animal Behavior Processes*, 3, 144–155.

Cartwright, N. (1983). *How the Laws of Physics Lie*. 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., and Robbins, S. J. (1990). Cue interaction in human contingency judgment. *Memory & Cognition*, 18, 537–545.

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

Cheng, P. W., and Novick, L. R. (1991). Causes versus enabling conditions. *Cognition*, 40.

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

Cheng, P. W., and Novick, L. R. (1995). From covariation to causation: A causal power theory. Unpublished manuscript, Department of Psychology, University of California, Los Angeles.

Estes, W. K., Campbell, J. A., Hatsopoulos, N., and Hurwitz, J. B. (1989). Base-rate effects in category learning: A comparison of parallel network and memory storage-retrieval models. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 15, 556–571.

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

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

Holland, J. H., Holyoak, K. J., Nisbett, R. E., and Thagard, P. (1986). *Induction: Processes of Inference, Learning, and Discovery*. Cambridge, MA: The MIT Press.

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

Kaplan, P. S., and Hearst, E. (1985). Excitation, inhibition, and context: Studies of extinction and reinstatement. In P. D. Balsam and A. Tomie (Eds.), *Context and Learning*. Hillsdale, NJ: Erlbaum. Pp. 195–224.

Kasprow, W. J., Schachtman, T. R., and Miller, R. R. (1987). The comparator hypothesis of conditioned response generation: Manifest conditioned excitation and inhibition as a function of relative excitatory strengths of CS and conditioning context at the time of testing. *Journal of Experimental Psychology: Animal Behavior Processes*, 13, 395–406.

Kaufman, M. A. and Bolles, R. C. (1981). A nonassociative aspect of overshadowing. *Bulletin of the Psychonomic Society*, 18, 318–320.

Kelley, H. H. (1967). Attribution theory in social psychology. In D. Levine (Ed.), *Nebraska Symposium on Motivation, 15* Lincoln: University of Nebraska Press. Pp. 192–238.

Koslowski, B., Okagaki, L., Lorenz, C., and Umbach, D. (1989). When is covariation not enough: The role of causal mechanism, sampling method, and sample size in causal reasoning. *Child Development*, 60, 1316–1327.

Lien, Y., and Cheng, P. W. (1992). How do people judge whether a regularity is causal? Paper presented at the 33rd Annual Meeting of the Psychonomic Society, St. Louis.

Matzel, L. D., Schachtman, T. R., and Miller, R. R. (1985). Recovery of an overshadowed association achieved by extinction of the overshadowing stimulus. *Learning and Motivation*, 16, 398–412.

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

Miller, R. R., and Matzel, L. D. (1988). The comparator hypothesis: A response rule for the expression of associations. In G. H. Bower (Ed.), *The Psychology of Learning and Motivation, Vol. 22*. San Diego, CA: Academic Press. Pp. 51–92.

Miller, R. R., and Schachtman, T. R. (1985). Conditioning context as an associative baseline: Implications for response generation and the nature of conditioned inhibition. In R. R. Miller and N. E. Spear (Eds.), *Information Processing in Animals: Conditioned Inhibition*. Hillsdale, NJ: Erlbaum. Pp. 51–88.

Nisbett, R. E., Krantz, D. H., Jepson, D., and Kunda, Z. (1983). The use of statistical heuristics in everyday inductive reasoning. *Psychological Review*, 90, 339–363.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems*. San Mateo, CA: Morgan Kaufmann.

Peterson, C. R., and Beach, L. R. (1967). Man as an intuitive statistician. *Psychological Bulletin*, 68, 29–46.

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

Rescorla, R. A. (1968). Probability of shock in the presence and absence of CS in fear conditioning. *Journal of Comparative and Physiological Psychology*, 66, 1–5.

Rescorla, R. A. (1972). Informational variables in Pavlovian conditioning. In G. H. Bower and J. T. Spence (Eds.), *The Psychology of Learning and Motivation, Vol. 6*. New York: Academic Press. Pp. 1–46.

Rescorla, R. A. (1981). Within-signal learning in autoshaping. *Animal Learning and Behavior*, 9, 245–252.

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

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. (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., and Dickinson, A. (1987). Associative accounts of causality judgment. In G. H. Bower (Ed.), *The Psychology of Learning and Motivation, Vol. 21*. San Diego, CA: Academic Press. Pp. 229–261.

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

Suppes, P. (1970). *A Probabilistic Theory of Causality*. Amsterdam: North Holland.

Suppes, P. (1984). *Probabilistic Metaphysics*. Oxford: Basil Blackwell.

Sutton, R. S., and Barto, A. G. (1981). Toward a modern theory of adaptive networks: Expectation and prediction. *Psychological Review*, 88, 135–170.

Wagner, A. R., Logan, F. A., Haberlandt, K., and Price, T. (1968). Stimulus selection in animal discrimination learning. *Journal of Experimental Psychology*, 76, 171–180.

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

Wasserman, E. A. (1990). Attribution of causality to common and distinctive elements of compound stimuli. *Psychological Science*, 1, 298–302.

Widrow, B., and Hoff, M. E. (1960). Adaptive switching circuits. *Institute of Radio Engineers, Western Electronic Show and Convention*, Part 4, 96–104.

Zimmer-Hart, C. L., and Rescorla, R. A. (1974). Extinction of Pavlovian conditioned inhibition. *Journal of Comparative and Physiological Psychology*, 86, 837–845.

IV Memory and Attention