©1992 The Institute of Mind and Behavior, Inc.
The Journal of Mind and Behavior
Winter 1992, Volume 13, Number 1
Pages 1–28
ISSN 0271–0137

Causal Knowledge: What Can Psychology Teach Philosophers?

Evan Fales and Edward A. Wasserman

The University of Iowa

Theories of how organisms learn about cause-effect relations have a history dating back at least to the associationist/mechanistic hypothesis of David Hume. Some contemporary theories of causal learning are descendants of Hume's mechanistic models of conditioning, but others impute principled, rule-based reasoning. Since even primitive animals are conditionable, it is clear that there are built-in mechanical algorithms that respond to cause/effect relations. The evidence suggests that humans retain the use of such algorithms, which are surely adaptive when causal judgments must be rapidly made. But we know very little about what these algorithms are and about when and with what ratiocinative procedures they are sometimes replaced. Nor do we know how the concept of causation originates in humans. To clarify some of these issues, this paper surveys the literature and explores the behavioral predictions made by two contrasting theories of causal learning: the mechanical Rescorla-Wagner model and the sophisticated reasoning codified in Bayes' Theorem.

David Hume, famously, gave us an analysis of the concept of causation which includes, in one of its versions, reference to a hypothesis about the origin of that concept. Hume had no way of empirically testing his hypothesis; whatever introspective grounds there are for his claim that experienced regularities produce expectations, one cannot move from this fact alone to the conclusion that expectations are the origin of part of our idea of a cause. Nor can anyone reliably remember the formation of this fundamental idea in his or her own case. But Hume was quite right, we believe, in seeing the relevance of the psychological question to epistemology.

In fact, there are several significant questions, among them: How do human beings (and other animals) apprehend the causal structure of the

This paper was supported by an Interdisciplinary Research Grant from the Center for Advanced Studies, The University of Iowa. Requests for reprints should be sent to Evan Fales, Department of Philosophy, or E.A. Wasserman, Department of Psychology, The University of Iowa, Iowa City, Iowa 52242.

world? As we learn to make causal judgments, what are the earliest or most primitive of these, and what are their sources in our experience? By what stages do strategies which yield causal understanding progress in complexity, culminating ultimately in the most sophisticated procedures of scientific reasoning? Does causal reasoning ever make use of some real or imagined constituent of necessary connection; and if so, what is its source and what use is made of it? Finally, are causal hypotheses arrived at and confirmed on the basis of single experiences, by means of enumerative induction, through Bayesian reasoning, or by some other means?

These questions are all of them empirical ones, and most, if they can be answered at all, must be settled by psychologists, not by philosophers; yet, they should be of great interest to philosophers, given the central place of causal relations in "all reasonings concerning matters of fact," as Hume puts it. Nevertheless, philosophers have paid rather little attention to these questions in recent decades (which contrasts with the early years of this century); and even Hume was surprisingly cursory in that survey of his own experience, which failed to find any objective constituents of causal relations beyond spatiotemporal contiguity and constant conjunction. At the same time, the last 40 years have seen the emergence of the empirical study of causal thinking in humans and animals by psychologists.

Perhaps the importance of the questions we have raised needs no justification for those epistemologists who look to science to provide an account of the acquisition of knowledge. Those who claim that evolutionary selection guarantees the reliability of sensory processes are also prone to think that evolution has secured the ability of organisms to determine causes. But this is too swift. Clearly, fitness only requires of an organism that it behaves roughly as if it knew how to "save the phenomena"—those phenomena, in particular, critical for its survival.

But the empirical findings, even those which go beyond introspection, ought also to interest those philosophers who insist that epistemology must begin with the egocentric predicament. Foundationalists, for example, should ultimately be concerned to show how our empirical evidence serves to justify those scientific theories that best explain the processes by means of which that very evidence is acquired.

The purpose of this essay, then, is to survey a number of different approaches which can be and have been taken in investigating the formation of causal judgments. It will become clear that there is a great deal we do not now know about these matters; thus, we shall raise more questions than we answer. Still, we hope in this way to bring to the attention of philosophers the progress that has been made, and the nature of some of the investigations currently underway. We shall also take occasion to propose some avenues for future research.

The Basis in Experience for Causal Judgments

Experimental research on causal judgment was largely initiated by the work of Michotte (1946/1963) on causal perception. Michotte performed a series of experiments in which adult subjects were presented with colored shapes whose motions could be analogized more or less closely to various collision phenomena; parameters such as velocity and direction of motion, spatial contact, and lapse of time between the stopping of one shape and the onset of motion in another could be controlled by the experimenter. When these kinematic variables closely reproduced those of common dynamic processes, subjects spontaneously reported what they saw in causal terms (e.g., that they saw the red square make the green one go forward). Moreover, only a single experience was sometimes required to prompt such reports, whereas no number of repetitions elicited causal descriptions when the parameters deviated from those which mimicked collision processes. Michotte regarded these observations as evidence that subjects were able to make noninferential causal judgments on the basis of visual data. He speculated that the basis of these judgments is the perception of a transfer of motion, possessing a kind of genidentity of its own and distinct from the moving objects, from one object to another.

Among philosophers, Harré and Madden (1975) have given prominence to Michotte's results, which they see as confirming their analysis of causation in terms of the action of "powerful particulars." Similarly, Michotte's work might be taken to support singularist theories of causation (Anscombe, 1971).

Piaget (1971/1974), on the other hand, criticizes Michotte's conception on the ground that the transfer of a quantity of motion—which retains its identity through the transfer and thus links the motions of two objects—is not something that can be *seen*. Perception of a causal connection between mover and moved, therefore, involves a construct. And, indeed, if Piaget's phenomenological claim is correct, it will not be difficult for a Humean to deflect the implication of Michotte's results, even allowing for the spontaneity of subjects' causal judgments. For the spontaneity may result from what Hume calls habit, previously established. The fact that Michotte's subjects are adults suggests that long prior experience has firmly entrenched the inclination to interpret certain of Michotte's displays in causal terms, and has produced resistance to seeing others in that way.[1]

[1] Subsequent studies of Michotte's phenomena have some bearing on the Humean argument, but they are not decisive. Yéla (1952) found that Michotte's results could be essentially duplicated when projectile A stops short of a target object B and B directly begins to move. Yéla argues that this runs counter to natural experience, especially since motion transferred across the gap must be "instantaneous" to produce a strong causal impression. But Yéla forgets that many previously experienced forces appear to be transmitted instantly across empty space—e.g., mag-

Piaget's own work—which attempts to account for the development of causal understanding in children—derives some inspiration from the much earlier work (ca. 1800) of Maine de Biran. Biran (1942) held that the source of the idea of causation lay in the subjective impression of unity between the subject and his/her actions, a unity which incorporates willing and muscular effort. Hume, of course, considers and rather summarily dismisses volition and effort as sources of the idea of causation. Piaget (1930/1966; see also Uzgiris, 1984) adopts Biran's notion that the subjective experience of force is the source of causal cognition. In his later work, Piaget (1971/1974) emphasizes the interplay between the development of children's conceptions of logical "operations" and their understanding of causal relations. In visually perceived collisions, a child must construct causal connections, and this construction would not be in any case possible if the child were unable to refer to and generalize from tactual and kinaesthetic data.

According to Piaget, the constructions of logical and causal conceptions proceed hand in glove through a series of stages which can be traced from infancy to pre-adolescence; but most of this development takes place after the acquisition of language. Experiments which allow children to manifest causal reasoning through nonverbal behavior now suggest that much causal reasoning occurs at a far earlier age than Piaget suspected (see Golinkoff, Harding, Carlson, and Sexton's 1984 review; also Berzonsky, 1971; Bullock and Gelman, 1979; Bullock, Gelman, and Baillargeon, 1982; Elek, 1990; Sedlak and Kurtz, 1981; Shultz, 1980). One striking observation which confirms this point has been reported by Watson and Ramey (1972), who showed that eight-week-old infants regularly learn to control the temporary motion of a mobile by head movements on a pillow containing a pressure-sensitive switch. Watson and Ramey additionally noted that infants who control a mobile smile more frequently at it than do infants who have just as frequent exposure to a turning mobile, but who have no control over its actions. Infants who had control and then were denied it sometimes displayed distress, even though still given exposure to a turning mobile. These observations suggest that an eight-week-old infant understands the difference

netic forces, gravity, and the like. Several studies suggest the influence of previous experience upon "spontaneous" causal judgments—e.g., Grüber, Fink, and Damm (1957), Gemelli and Capellina (1958), Powesland (1959), and Olum (1956, 1958). Olum tested 7-year-olds with a Michotte apparatus, but tried to explain the difference she found between adult judgments and theirs in terms of developmental differences in the perception of purely kinematic features. More interesting results are recorded by Leslie (1982), who habituated $4^1/_2$- to 8-month-old infants to Michotte-type motion pictures of "normal" and time-delayed collisons, and then recorded their degree of attention (judged by eye movements) when shown movies in which the target projectile fails to move after being struck or moves without being struck. The results suggest that these infants spontaneously attend to at least the feature of continuity of motion. Whether this should be interpreted in Michotte's fashion remains an open question.

between the events it causes and those over which it does not exert control. We turn now to some questions raised by such results.

The Primitive Sources of Causal Judgment

It would be of great interest to know whether all of our capacity to judge causes and effects can be traced back to experiences of a single sort; or whether experiences of a variety of sorts can "trigger" such judgments. However, because the acquisition of causal understanding seems to begin in the earliest stages of preverbal infancy (if not before birth), it is a difficult process to investigate. For example, one of us has argued elsewhere (Fales, 1990) that the most primitive and essential experience which yields a concept of causal relation is the tactile and kinaesthetic experience of pushes and pulls, whether or not accompanied by volition. This view has it that the idea of a necessary connection between events—which is a constituent of our idea of causal connection—derives from sensations of force. One way to draw the contrast between such a theory and Hume's classical analysis of causation is to recognize that, on Hume's theory, the result of a felt push or pull could not in any sense be predicted on the basis of a single experience, whereas it could be predicted, in principle, on the alternative view. Such an alternative would thus take tactual recognition of causal relations to be primary; inductive reasoning would be a secondary and parasitic process which would take over where there was no experience of felt force, but only of repeated patterns.

Now human subjects, no matter how young at the time of testing, will already have had repeated tactual experience of force and its consequences; and tactual sensitivity to contact forces is present in all animal organisms except perhaps the most primitive. Thus, the circumstances required for the envisioned experiment unfortunately cannot be satisfied except in imagination; though some indirect evidence exists,[2] we cannot hope for a direct experimental resolution of the issue. But, indeed, let us suppose that such an experiment could be performed.

[2] Bullock, Gelman, and Baillargeon (1982) do report that 3- and 4-year-old children rely less on covariation in making causal judgments and more on spatial and temporal contiguity; also, in reasoning that one event caused another, they rely on the existence of an observed or imputed sequence of mechanical contacts linking the two. Shultz (1982) reports an ingenious set of experiments, some with children in Mali who had no previous exposure to the gadgets (e.g., tuning forks, flashlights) used by the experimenter. The experiments were designed so that causal judgments based on spatiotemporal contiguity and covariation were made to compete with judgments arrived at by positing some transmission of force. According to Shultz, the results show that even 3-year-olds favor reasoning of the second sort. But, unfortunately, it is unclear how much these children may have relied on analogies to previous experience. These data, while suggestive, in no way decisively favor the non-Humean hypothesis being discussed.

A subject who has not previously had any tactile or kinaesthetic sensations (perhaps because of a reparable neurological defect) now suddenly acquires them; experiences a single push on some part of the body and its subsequent motion; and is asked to form an opinion, on the basis of the single experience, about the outcome of a similar future push. Suppose the subject were to form the "correct" expectation. Although Hume's theory allows only the gradual formation of expectations by degrees, there is a way of responding to this hypothetical result which preserves Hume's more fundamental claim that no natural necessity is perceived to link pushing to motion. This response explains that result by postulating that the mechanism that associates the idea (or expectation) of an effect with a cause may in certain sorts of cases be triggered by a single experience to increase abruptly to full strength, rather than arising gradually through repeated experience. Whether there are such mechanisms is a question which must be resolved empirically; we mention this possibility because the existence of cognitive processes which associate causes with effects in a more or less mechanical fashion (whether abruptly or gradually) is plausibly predicted by evolutionary theory. It serves an organism well to be sensitive to the causal contingencies of its environment; and because of the great survival value of such sensitivity, we might reasonably expect it to evolve in primitive forms of life which do not think at all, and to be mediated in them by quite simple mechanisms. When we reflect that any kind of conditioning poses the problem of recognizing a causal contingency, it becomes apparent that very primitive forms indeed are capable of some learning of causal relations (at least in the sense of developing systematic and appropriate responses). For example, *Hermissenda*, a nudibranch mollusk, can be conditioned to associate light with rotation, and the strength of this conditioning is degraded when light signals unpaired with rotation or rotations unpaired with light are inserted into the training schedule (see Farley, 1987a, 1987b).

Animal conditioning experiments, on which there is an extensive literature (see Kimble, 1961; Mackintosh, 1974), largely investigate the capacity to respond to causal contingencies, where the evidence for those contingencies takes a "Humean" form; that is, the animal is presented with certain patterns of covariation between observable events, but the events are not directly linked as experienced pushes to experienced motion, and the mechanisms which in fact link event-pairs are not observable by the animal. Hence, even if causal relations are in other circumstances directly perceivable, the only recourse available here is induction, or something mimicking it. This literature is surveyed in the next section, where we consider what is known about factors which control the ability to learn causal relations under these conditions.

That discussion will shed some light on a fundamental issue which must now be raised. We have just alluded to the fact that even very simple forms of animal life exhibit a capacity to acquire differential and plastic responses to causal contingencies. Placing this behavior at one extreme, we may put at the other the sophisticated use of probability computations and statistical methods devised to determine causal relations by some professional gamblers and by scientists. In the latter case, there is an explicit and well-articulated use of mathematical theory to provide the quantitative computational strategies sometimes required to reveal subtle causal relations.

Clearly, it will not do to explain the causal sensitivities of *Hermissenda*, say, by ascribing to them the capacity for propositional thought or mathematical calculation. Whether or not we ascribe some primitive sort of awareness to these mollusks, we will want to show how their learning abilities are mediated by simple algorithms embodied in essentially mechanical processes of some sort, as Farley (1987b) has begun to show. Let us say that in *Hermissenda*, we expect a mechanistic explanation of response to causal contingencies, whereas in scientists and clever gamblers, it is reason which supplies the conclusions. (This is surely a viable distinction, even if it could be shown that rational thought is embodied in neural mechanisms, and even if it could be shown that there is a continuum between the procedures of scientific method and the primitive algorithms used by *Hermissenda*.) Now our question may be crudely put as follows: By what stages, and how, is the transition made between mechanistic learning and reasoned cognition?

Mechanism vs. Reason

One of the earliest mechanistic theories of learning (of causal relations) was put forward by Hume in his *Treatise*. Having despaired of solving the problem of induction, and hence of giving an account of learning which shows how our causal judgments can be founded upon adequate *reasons*, Hume (1758/1955, pp. 55–56) proposes that the causal judgments we anyway constantly make must be generated in an essentially mechanistic way, by processes with which nature has fortunately provided us, and over whose operation we have little rational control. Hume's proposal is that the regular recurrence of a pattern of spatiotemporally contiguous events produces in the mind a progressive association of the "idea" of the first sort of event with that of the second—an association which results in a gradual increase in the strength with which, given an event of the first sort, we expect the subsequent occurrence of an event of the second. It is noteworthy that Hume supposed his model to be applicable to human beings and to other animals alike.

Because the elements of Hume's model are mental items, such as ideas and sensory impressions, it might seem inappropriate to characterize his theory as mechanistic. However, we believe the term is entirely apposite: the intended contrast here is between the workings of reason or understanding and the workings of a machine. The way in which ideas come to be associated (or disassociated) is taken by Hume to be governed by laws in much the same way as the interaction of physical particles might be.

As pointed out above, Hume's theory is not the only mechanistic theory of learning one might entertain. Thus, one might imagine that some associations are innate, in the sense that once the "ideas" arise from sensory impressions, they are automatically and immediately associated in ways which do not depend upon regular conjunction. Or, one might hold that every idea, once formed, has a tendency to be associated with every other, the function of experience being progressively to disassociate those ideas which do not correspond to impressions regularly conjoined. Nor are these alternatives the only imaginable ones. But in any case, it is Hume's model, or refinements of it, which have come to be adopted by many contemporary psychologists, and which seem indeed to be best confirmed by the experimental data on animals and humans.

Here, we shall survey some of these learning theories and the experimental results which bear on them. We shall also propose a rational model of causal judgment formation, which largely duplicates the predictions of the mechanistic models, and ask whether there are divergent predictions which would permit an experimental decision between them. The rational model we propose is well known to philosophers, but it has not, so far as we are aware, been proposed or tested empirically by psychologists. We shall then return finally to the question posed at the end of the previous section: Where does mechanism leave off and reasoned understanding begin?

Experimental Findings

According to Hume, three conditions must be met if we are to say that an event, C, is the proximate cause of another, E: (1) C must immediately precede E, (2) C must be spatially and temporally contiguous to E, and (3) C and E must be of types which are constantly conjoined. Although Hume does not allow causal connections which are probabilistic, he does allow that causation may be "mixed with" chance, and also that events may be partially brought about through the action of hidden causes. Thus, the observed covariation between C and E, upon which we base a judgment of causal connection, may be statistical. Here, the strength of our judgment is a function of the strength of the covariation. (Hume does not clearly distinguish causal strength from the strength of our belief that the causal relation holds.)

Psychologists have attempted to discover the relative dominance of these three factors, among others, in the production of causal judgments. Three other factors in particular are worth mentioning: the perceptual salience of the stimuli (C and E); the strength of reinforcement, where one of the events (typically E) is a reinforcer; and, where there is a temporal or spatial gap between C and E, the existence of other cues which might facilitate the construction of a chain of events leading from C to E.

Strength of conditioning is usually degraded by the introduction of a temporal gap between cause and effect. Animal studies show that in most cases,[3] a gap of several seconds is enough to affect learning significantly. Experiments by Mendelson and Shultz (1976), Siegler and Liebert (1974), and Siegler (1975) with four- to nine-year-old children suggest that, especially in children younger than eight years, delays of only five seconds inhibit causal judgments (see also commentary in Sedlak and Kurtz, 1981). Cues which permit an explanation of the delay in terms of an intervening causal process significantly restore causal judgments; but failing this, temporal contiguity seems to take precedence over invariable covariation in supporting causal judgments. Furthermore, temporal precedence appears to outweigh spatial contiguity and to be relied upon by children as young as three years in making causal judgments (see Bullock and Gelman, 1979; Sedlak and Kurtz, 1981).

A considerable body of data has been collected on the ability of humans and animals to assess causal connections, data which measure the strength of association between a conditioned stimulus (CS) C which is followed by an unconditioned stimulus (US) E. The parameters which have been varied in such experiments include: the frequency with which C is not followed by E, the frequency with which E is not preceded by C, the complexity of C and E, and the addition or subtraction of stimuli to or from C and E. E is usually a positive reinforcer or punisher, e.g., food or electric shock.

Hume's discussion of causation evokes a simple picture of the conditions under which causal beliefs are acquired: stimulus C is invariably followed by stimulus E;[4] E immediately follows C and occurs next to it or in the same place; and C–E pairings are spaced well apart, so as not to confuse the issue of which C pairs with which E. However, even lower animals are able to establish causal connections where there is a temporal gap between a C and its paired E; where the gap is not of like duration from one pairing to the next; where there are (variable) spatial gaps; where unpaired occurrences of C or E intervene; and where sequential Cs and Es occur so closely spaced that it is unclear which E (if any) pairs with which C. To be sure, such complexities degrade both speed of learning and the strength of the acquired belief. Gormezano and Kehoe (1981) discuss these data.

[3] There are notable exceptions. Rats, for example, will learn to avoid foods that make them ill after only a single aversive experience, even when the onset of illness occurs some hours after the ingestion of food. Psychologists disagree over whether in such cases generalized learning mechanisms are superseded by genetically coded specialized sensitivities important for survival (see Domjan, 1983).

[4] Although Hume (1739/1888, pp. 125–137), explicitly allows for the case where E sometimes fails to follow C, the covariation need only exceed chance.

The effect of temporal relations between paired Cs and Es is actually quite complex and not yet well understood. However, certain general features emerge from the experiments which have been so far performed. Supposing the duration of events C and E to be relatively short, we would expect that pairings would be apparent where the interval between a C and its paired E—the interstimulus interval or ISI—was short compared to the interval between one presentation of C and the next one (the intertrial interval or ITI). And this is what we find: when the ISI/ITI ratio is small, learning is much more rapid than when it approaches or exceeds 1. Alternatively, one can use Cs which last until, or past, the onset of an E event, and measure learning rates for short-duration Cs versus long-duration Cs. Here, C and E are contiguous, but the *onset* of C is variably related to that of E. Again, learning is markedly greater when the C-onset to E-onset time is small compared to the ITI. However, the data also suggest that the onset of a stimulus is more salient than its continuation; thus, a long C-onset to E-onset time functions, as does a long ISI, to degrade learning even where these times are small as compared to the ITI. (For a discussion of the data and the literature, see Gormezano and Kehoe, 1981; also Gibbon and Balsam, 1981; Gibbon, 1981; and Jenkins, Barnes, and Barrera, 1981. Could there be a delay between the onset of a cause and its immediate effects? On this question, philosophical opinion is divided. See e.g., Hume 1739/1888, pp. 75–76; Russell 1917, pp. 136–140.)

None of this will be surprising. More interesting, perhaps, are data which, although equivocal, suggest that contingency judgments are weakened when paired Cs and Es occur simultaneously rather than sequentially, in animals (see Gormezano and Kehoe, 1981). This result might be taken to suggest that causal connections are discounted, relatively, where one event is not perceived to precede another. Conceivably, this is because the observed contingency could be explained by an unobserved common cause for both events; but, alternatively, it may be that one of the stimuli draws attention away from the other with which it coincides.

In any case, organisms do learn even when ISI/ITI ratios are large and when unpaired Cs and Es are presented. So one must ask: How do they manage it? To see that there are real complications here, let us consider three strategies by means of which a subject might plausibly arrive at a causal association between C and E, where there are temporal gaps and instances of unpaired Cs or Es. If we designate by ~C those times when no C is occurring, we have the following strategies:

I. Measure the probabilities (frequencies) with which E follows C (=P(C/E)) and with which E follows the absence of C (=P(E/~C)), and take their difference as a measure of the causal effectiveness of C in producing E:

$$\mathcal{E}_c = \Delta P = P(E/C) - P(E/{\sim}C).$$

II. Measure the average time interval T(E/C) between any C and the next occurrence of E, and compare this to the average time interval (T(E/~C) between a nonoccurrence of C and the next E. If T(E/C) < T(E/~C), then C causally facilitates E.

III. Measure the average time between successive occurrences of E, $T(E_{i+1}/E_i)$. Compare this to the average time between occurrences of C and the next succeeding occurrence of E, T(E/C). A positive value for $T(E_{i+1}/E_i)$–T(E/C) indicates that C causally facilitates E.

Now under certain circumstances, at least, these measures of C's "strength" in producing E are not only qualitatively, but quantitatively equivalent. Imagine a training period, for example, which consists of successive one-second long "runs." An E either occurs or fails to occur at the end of each run; if E does occur and a C (one or more) occurs at any time during that run, this scores as one positive instance of C–E co-occurrence; if C occurs without E or vice versa, a negative instance is scored. Under these circumstances, it is easy to show that T(E/C) is inversely proportional to P(E/C); and similarly for T(E/~C) and P(E/~C); thus, ΔP is proportional to 1/T(E/C)–1/T(E/~C). This congruence does not, however, entail that the ΔP measure and the time-based measure have equal "psychological reality" for subjects. (For an experimental investigation of various metrics to predict human perception of causal relations, see Chatlosh, Neunaber, and Wasserman, 1985; Wasserman, Chatlosh, and Neunaber, 1983. ΔP is the most successful metric for predicting asymptotic judgments of causal strength.)

If, however, the subject does not know the just-described details of the experiment, then the following puzzles arise with respect to the possibility of making either relative frequency or average time-lapse estimates:

> a) If Es, when they follow Cs, do so after a time lapse—i.e., if Cs occur at the option of the subject but Es only at the end of a trial—then why don't these time lapses count as occurrences of ~C and preempt the covariation between Cs and Es?
>
> b) If there are such time lapses between Cs and Es, how is the subject to divine that a C is to be correlated only with an E which follows it by less than one second? What is to prevent C–E pairings from overlapping? Indeed, a *variable* time-lapse would seem to encourage such speculation.
>
> c) Clearly, estimates of frequency and of time lapse will depend upon the principles by means of which the subject individuates events of type C, ~C, E, and ~E. So if C does not occur during a one-second trial, to how many occurrences of ~C does that correspond?

Even in the artificially simple situation envisioned here, a naive subject has no objective basis for answering these questions. In real life, where multiple factors may conspire to produce an event, where time gaps are variable, where irrelevant and distracting events are numerous and relevant pairings may occur infrequently or so quickly as to overlap, the problem begins to assume a formidable complexity. Nevertheless, since subjects—even animals—can learn correlations of the sort described, sometimes with amazing rapidity and accuracy, we must ask what factors other than co-occurrence might facilitate such learning. It may be that in natural settings, spatial regularities between cause and effect, qualitative similarity between them, and covariation of strength between variable-strength stimuli, can also serve as cues (see Watson, 1984).

Is there any "natural" period of time after the occurrence of a CS event during which an animal is on the alert for correlates with it, and after which succeeding events will not be associated with it? What if more than one CS event occurs during such an interval or more than one potential US event occurs with which it (they) may be correlated? We remarked above that degradation of conditioning is produced by CS–US time lapses, to a degree which is dependent upon the species and upon the nature of the CS and US. This fact places an upper bound on the temporal interval which can obtain between a CS and a US which are subjectively perceived as paired, but it does not solve the "overlap" problem within that interval.

At the other extreme, we might ascribe to subjects a model of causal connection which requires *proximate* causes to precede *immediately* their effects without gap. On this hypothesis, causal judgments connecting temporally noncontiguous events will occur only where some feature(s) of the situation allow(s) the subject to interpolate some hypothetical chain of connecting causes linking the events. But then it becomes incumbent upon psychologists to specify those features and to explain how they might operate to suggest such hypotheses (see Einhorn and Hogarth, 1986, for a brief discussion). Moreover, although it is not implausible to impute such (perhaps very vague) hypotheses to human subjects, it is more problematic whether animals such as rats reason in this way. (We shall return shortly to the issue of how intellectual the formation of causal judgments is.)

With respect to the matter of individuating events, we might plausibly take our cue from C.J. Ducasse's (1951) analysis of causation, according to which the cause of an event X is the total *change* in the spatially contiguous circumstances which immediately precedes it. The defects of Ducasse's account as a philosophical analysis are well known; but this is no bar to its usefulness as a psychological suggestion. It may be that only perceptually salient changes are attended to and regarded as potential causes of subsequent changes; from this perspective, if A represents a change which occurs against relatively stable "background," an X which follows an A closely enough will be paired with it, and unpaired Xs will be regarded as uncaused, or as having hidden causes. Taken in conjunction with a specification of the maximum allowed temporal gap, this assumption would help us to answer questions (a), (b), and (c), posed four paragraphs previously.

Learning Causal Contingencies: Mechanistic Theories

We turn now to a major theory of conditioning and summarize some of its relevant aspects. The theory is the contiguity model of Rescorla and Wagner (1972). It will seem to be reductionistic in the sense that it attempts to explain the learning process in quasi-mechanistic and molecular terms (that is, in terms of the cumulative effect upon an organism of the moment-by-moment impingement of stimuli upon it). We shall contrast this theory with a normative and intellectualistic model of learning more familiar to philosophers, *viz*. Bayes' theory of confirmation, applied in this instance to causal hypotheses.[5] Having introduced these models of the learning process, we will ask whether there are empirical results which do or could determine which one of them gives a more nearly correct account of the process by which causal beliefs are formed.

Rescorla and Wagner (1972) present a theory of classical or Pavlovian conditioning which is mechanistic—it relies upon laws which govern the formation of associative strengths between stimuli—and molecular; that is, associative strengths are modified in step-by-step fashion as new experiences of stimuli are registered.[6] According to Rescorla and Wagner, every US which acts as a reinforcer or punisher (e.g., food or electric shock) supports a certain maximum degree of associative strength for an organism. This strength is a measure of the vigor and reliability of those behaviors by means of which the organism shows that it "expects" the US when presented with the CS, in the limit where further training does not achieve any further increase in learning. Given a series of CS–US pairings, an organism will learn to associate the two. Calling the CS 'A,' we use 'V_A' to designate the strength of association to the US which exists at a given time, and 'λ' to designate the maximum associative strength that the US will support. If we use behavior as a measure,[7] we find that organisms learn gradually to associate a

[5] J.S. Mill's rules of induction (1843/1911) can be regarded for our purposes as subsumed by Bayesian reasoning. A spectrum of other models of learning has been proposed, ranging from intellectualistic ones which posit the operation of rational principles, to mechanistic ones which depend on laws of association or refer to neurophysiological processes. A good overview is provided by Gormezano and Kehoe (1981).

[6] Thus, an organism need not remember the particulars of its previous history of relevant experiences, as it would if it were performing an inductive inference using the straight rule, a rule whose application requires keeping a score sheet of positive and negative instances of a regularity.

[7] If by learning we mean something conceptual (viz. coming to recognize a connection between the CS and US), then this use of behavior requires a number of assumptions, since there is no a priori guarantee that strength of belief (or whatever corresponds to it in animals) maps linearly onto strength of behavior. For example, an animal may compensate for the intrinsic cost (e.g., expenditure of effort) of the behavior, or it may lose or gain interest in acting on its belief, or lose or gain interest in the reinforcer itself. But we will assume that these difficulties can be discounted—a not unrealistic assumption in many cases.

CS with a US: changes in associative strength are initially large and then decrease to zero as learning goes to asymptote. A key idea in Rescorla and Wagner's approach is the claim that a change in associative strength is a linear function of the difference between the maximum associative strength, λ, allowed by the US, and the associative strength, V_A, which has already accrued to it:

$$\Delta V_A = \alpha_A \beta(\lambda - V_A)$$

Here, α_A is a learning rate parameter which depends upon the CS, and β is a learning rate parameter which depends upon the US.

If we assume that lack of a reinforcing US no longer supports an association (rather than, for example, supporting a negative association) between the CS and the US, then we can account for the extinction of a learned response: now, V_A decreases at a rate proportional to its previously acquired strength. Hence,

$$\Delta V_A = \alpha_A \beta(0 - V_A) = -\alpha_A \beta V_A.$$

The main strength of the Rescorla–Wagner (R–W) theory lies in its ability to explain the results of a wide range of experiments involving compound and multiple CSs. Two puzzles of central importance are provided by the phenomena of blocking and overshadowing. In blocking, the degree to which a subject has learned to associate a CS, X, with the US is determined after training under each of the following two schedules: A^+; AX^+ *versus* AX^+. (Note the change of notation: 'A' and 'X' now stand for CSs, '+' for a US, and '–' for absence of a US. Here, A^+ indicates that the subject is trained to associate a CS [A] with a reinforcer [+], and the semicolon indicates that this is followed by the training sequence AX^+, in which two stimuli, A and X, are presented simultaneously, followed by the reinforcer.) The subjects of the control group, which are given only AX^+ training, are presented with the same number of Xs paired with the US as are subjects of the first group. Under these conditions, the second group learns to associate X with the reinforcer much more strongly; in the first group, the previously learned A^+ association apparently blocks conditioning to X.

To explain this result, the R–W theory introduces a second central postulate: when a stimulus is part of a compound CS, changes in its strength depend upon the total associative strength of the compound. Thus, under an X^+ schedule (or for that matter, an A^+; X^+ schedule), increments in the associative strength of X depend only upon the magnitude of the already-acquired associative strength of X: $\Delta V_X = \alpha_X \beta(\lambda - V_X)$. But under an AX^+ schedule, ΔV_X is determined by $\Delta V_X = \alpha_X \beta(\lambda - V_{AX})$ and ΔV_A by $\Delta V_A = \alpha_A \beta(\lambda - V_{AX})$. Now Rescorla and Wagner assume that the strength V_{AX} of a compound is simply the sum of the strengths of its components: $V_{AX} = V_A + V_X$. Given this, it is easy to see how blocking works. Under the schedule A^+; AX^+, V_A acquires a proportion of the total associative strength,

λ which the US will accommodate;[8] thus, the value of V_{AX} may already be high, relative to λ, at the start of exposure to AX^+. In that case, increments in V_X will be small; V_A has "gobbled up" most of the associative strength that the US will support. Intuitively, we might say that the subject discounts the relevance of X if A seems already sufficient to account for the occurrence of the US; but this rationalistic way of putting it would misrepresent the essentially mechanistic character of the R–W theory.

Overshadowing is exemplified where conditioning to a CS, X, presented alone, is compared with conditioning to that CS when presented in conjunction with a second CS, A. That is, we test for the strength of conditioning to X after the two schedules AX^+ *versus* X^+. Not surprisingly, we find that V_X is weaker after AX^+ than after X^+; A "overshadows" X. In the R–W theory, the learning-rate parameters α_A and α_X determine the extent to which A overshadows X. Hence, when A is paired with X, V_A gobbles up some proportion of the associative strength λ supported by the US, and V_X is correspondingly weakened.

We can summarize briefly some further data which the R–W theory handles nicely. The reader can easily supply the relevant explanations where we omit them.

(1) Consider the schedule A^+/A^-. (Here, 'A^-' indicates trials in which A is unreinforced, and the '/' indicates that unreinforced trials are interspersed randomly among reinforced trials.) This schedule results in weaker conditioning to A, depending upon the proportion of nonreinforced trials.

(2) The schedules AX^+/A^+, AX^+, and AX^+/A^- lead, respectively, to weak, moderate, and strong conditioning to X.

(3) Let A be a frequently occurring stimulus and B an infrequent one in an initial $A^+/B^+/X^+$ presentation. Then the schedule $A^+/B^+/X^+$; AX^-; X^+ leads to slower reacquisition of conditioning to X than does $A^+/B^+/X^+$; BX^-; X^+. (Intuitively, the AX^- sequence marks X as a more potent inhibitor than does the BX^- sequence, since A is a stimulus with greater associative strength than B.)

(4) When USs unsignalled by a CS are interspersed among CS–US pairs (i.e., $A^+/{\sim}A^+$), conditioning to the CS is impaired compared to the case in which no unsignalled USs are given. (Rescorla and Wagner explain this finding by hypothesizing that the "constant" background—call it 'B'— acts as a CS, so that we can more accurately represent this schedule by $AB^+/B^+/B^-$; that is, B alone is sometimes reinforced and sometimes not. So long as the reinforcement rate for AB exceeds that for B alone, there will be some conditioning to A, but B will gobble up some of the associative strength which otherwise would accrue to A.)

We could extend this list of cases, but it should suffice to suggest the power of the R–W theory. Instead, we turn to some results which the theory in its stated form does not seem to be able to handle. Mackintosh, Dickinson, and

[8] The magnitude of V_A will depend upon the length of the training to A+.

Hall all report studies which investigate in greater detail the phenomena of blocking and overshadowing (see Dickinson, Hall, and Mackintosh, 1976; Dickinson and Mackintosh, 1979; and Mackintosh, 1971). Mackintosh (1971) performed an experiment in which he demonstrated the effect of overshadowing on the first trial of conditioning, i.e., before either component had acquired any response strength. The R–W theory predicts that a stronger stimulus should "squeeze out" a weaker one only as conditioning proceeds toward asymptote. Mackintosh (1971) also has results which suggest that pretraining to stimulus A can completely block conditioning to another stimulus X under exposure to AX^+; R–W predicts that V_X acquires at least some strength.

More interesting, perhaps, are the results of experiments which investigate whether the blocking effect is weakened by changes in the US which are designed to "surprise" the subject. For example, Dickinson et al. (1976) compared the strength of conditioning to stimulus X after the following two schedules: A^+; AX^+ *versus* A^+; AA^{++}, where the double '+' indicates an extra presentation of the reinforcer (electric shock) or a stronger reinforcer in the AX condition. The second schedule produced stronger conditioning to X. Similar results are obtained when a single US is used in both schedules, but the second group is surprised by postponement of the US for a few seconds under the AX^+ condition. R–W can handle the first unblocking result by arguing that the added US supports a higher associative strength; i.e., it increases the value of λ and hence provides more "room" for associative strength to accrue to V_X. But it has difficulty with the second result [Why should postponement of the shock increase the strength of association the US will support?] and even greater difficulty in accounting for the unblocking which results from schedule A^{++}; AX^+, in which omission of an expected US also improves conditioning to X.

Shanks (1987) has recently emphasized that the R–W theory gives qualitatively correct predictions of learning curves. To take a simple case, if we plot the degree of conditioning to a CS, A, which is invariably or partially reinforced as a function of the number of trials, we obtain a curve which rises quickly at first and then levels off asymptotically to some maximum. This is not what we should expect of an organism which utilizes the straight rule of enumerative induction, or metrics such as ΔP, which rely upon relative frequencies. If we assume that the organism is presented with a fair or random sampling upon which to base its estimates of relative frequency, and if we assume that the measured behavior (the conditioned response) is a linear function of the degree of covariation the organism assigns to CS–US pairings, then we should expect behavior (hence estimated degree of contingency) to show rather marked variation during early trials, but centered around the "true" or asymptotic value. With increasing trials, these oscillations will damp down to converge on that value.

Of course, we must distinguish the organism's judgment of the *strength* of the CS–US connection from the degree of *confidence* which it assigns to that judgment. Ideally, the latter will (if relative CS/US frequencies are used to estimate relative probabilities) increase in negatively accelerating fashion from zero to one; so perhaps subjects conflate contingency judgments with confidence estimates.

Shanks has investigated the latter possibility in human adults. He found that subjects could make judgments of the causal strength of a CS and offer separate estimates of the confidence they attached to those judgments. The confidence estimates increase as subjects acquire more data, even when those data lead them to rate causal strength at zero or lead asymptotically to a negative causal rating (the CS prevents the US). Thus, Shanks concludes that it is unlikely that such conflation occurs. But he seems to assume that conflation would lead subjects simply to sum their numerical estimates of causal strength and of confidence; he ignores, for example, the much more plausible supposition that subjects entertain a number of competing hypotheses about causal strength, assign them varying degrees of likelihood in the light of the data, and, forced to offer a single judgment of causal strength, make an estimate by averaging hypothesized causal strengths, weighted by their respective confidence ratings (which amounts to judging the total subjective probability that the next CS will be followed by a US; see Burkes, 1977, p. 77). This procedure would account for Shanks' results, as would two other strategies discussed below. All these strategies require, however, that subjects estimate the likelihoods, relative to evidence, of various hypotheses about causal strength; Bayes' theorem offers one formalization of how to make such estimates. Thus, whether the R–W model is in this important respect superior to inductive models (which posit running estimates of relative frequency as the strategy for judging strength of contingency or causal efficacy) remains to be seen. We turn, therefore, to a cognitive model of learning which employs Bayesian reasoning.[9]

Learning Causal Contingencies: Normative Theories

As we mentioned earlier, the Bayesian theory of confirmation provides a general model for reasoning about the strength of confirmation of a hypothesis; here, we apply it specifically to causal hypotheses. An essential way in which Bayesian reasoning differs from enumerative induction is this: enumerative induction evaluates the strength of a hypothesis relative to a given body

[9] Carnap (1952) demonstrates the existence of a continuum of such models; here, we restrict ourselves to the Bayesian model. For the theoretical development of another intellectualistic model, one which requires an organism to keep a running record of events and to calculate relative frequencies, see Granger and Schlimmer (1986).

of data without regard to other hypotheses which might also explain—and hence be supported by—the same data. A Bayesian, instead, always begins with a set of mutually exclusive (and in principle, exhaustive) hypotheses, which must compete for confirmation; if the hypotheses are competing and exhaustive, then we can set the sum of their epistemic probabilities equal to one. But then, each hypothesis must be assigned a prior probability; as the data come in, this probability will in general shift, so that the best-confirmed hypothesis at one point may be replaced by another at a later time.

This competition is in a sense analogous to the mechanism that is a central feature of the R–W theory—namely, competition among stimuli for associative strength. This parallel will be clarified if we show how the Bayesian approach predicts learning curves.

Let us suppose that the task is to estimate the causal strength of a CS in producing or preventing a US where, perhaps, the US does not always follow the CS and may sometimes occur without it. We may, somewhat artificially, imagine a subject to entertain 20 competing hypotheses. The twentieth, h_{20}, postulates a causal strength between +.90 and +1.00; that is, it predicts that the CS is strong enough to produce the US between 90 and 100% of the time, in the long run. The nineteenth hypothesis, h_{19}, assigns a causal-strength rating between .80 and .90, and so on. At the opposite pole, h_1 assigns the CS a strength between −.90 and −1.00; i.e., says that it *prevents* the US to this degree.

The general form of Bayes' theorem, which determines the degree, $P(h/e{\cdot}b)$, to which a hypothesis, h, is confirmed by the new evidence, e, given already existing background evidence, b, is:

$$P(h/e{\cdot}b) = \frac{P(h/b) \text{ x } P(e/h)}{P(e/b)}$$

Intuitively, the idea is that the subject makes running calculations of the values of $P(h_i/e{\cdot}b)$ for i between 1 and 20—these must sum to one—as each new piece of evidence (each paired or unpaired CS and US) is collected.[10] That evidence is then incorporated into the background when the next result is recorded; thus, the value $P_n(h_i/e{\cdot}b)$ of $P(h_i/e{\cdot}b)$ after the n^{th} trial is equal to $P_{n+1}(h_i/b)$, the value of $P(h_i/b)$ after the $n+1^{st}$ trial (see Burkes, 1977, pp. 65–92 for details).

$P(h_i/b)$, the prior probability of h_i relative to background information, must be assigned some initial value when no information whatever has yet been obtained, in order for the sequential calculations of its probability to

[10] A CS paired with a US tends, other things being equal, most strongly to confirm hypotheses with subscripts closer to 20 and to disconfirm those with low subscripts. A CS unpaired with a US has the opposite effect. A US unpreceded by a CS tends to raise the value of P(e/b), the probability that a US can occur "anyway."

proceed. Early on, the prior probabilities of the h_i's will dominate over the influence of the evidence; but the effect of evidence, as it accumulates, swamps that of the priors, and the hypothesis that wins the competition approaches asymptotically the one that predicts the asymptotic value of ΔP. Thus, Bayesian reasoning generates learning curves.

But what about the shape of these curves? We may note, first, that little empirical information exists as to curve shape very early in the learning process. (Shanks' data, which are typical, track learning only after the first 10 trials. See however Chatlosh, Neunaber, and Wasserman, 1985, especially pp. 13–15, for data on learning curves tracked from the first trial.) Yet, such information may be significantly indicative, e.g., of whether and how subjects assign prior possibilities to alternative hypotheses.

Now suppose we make the following assumptions. We interpret the learning curves as giving us straightforwardly a subject's best estimate of causal strength or ΔP—that is, as specifying the hypotheses thought most likely to be true. Then the predictions of the Bayesian model depend significantly upon the prior probabilities the subject assigns to alternative hypotheses (see also Alloy and Tabachnik, 1984). If, in our example, the subject applies a principle of indifference to assign each hypothesis a prior of 1/20, then the Bayesian model will not predict the observed learning curves: an initial CS–US pair will assign h_{20} the best probability, although only very slightly higher than any other hypothesis.

However, a Bayesian theory will yield learning curves similar to those predicted by the R–W theory if we modify these assumptions. We have already suggested that perhaps subjects' estimates of causal strength ought to be interpreted as weighted averages of those competing hypotheses they regard as "live" in view of the data. Alternatively, we must consider that subjects may *not* give every hypothesis equal prior weight. In particular, suppose subjects are biased toward an initial assumption of randomness or lack of causal connection between the elements of their experience, giving h_{10} and h_{11} the highest priors, with the values of the other priors tailing off to either side. This, too, will result in low initial ratings of causal efficacy, moving only gradually away from zero or near-zero values, if experience warrants. Such an "unfair" assignment of priors surely has considerations in its favor: organisms which jump to assigning high causal strengths on the basis of little data (even with low confidence) will tend to see causal connections everywhere, especially where experience does not afford frequent repetition (see Skinner, 1948); and they will be subject to frequent and often marked revisions of their causal beliefs.

There is yet a third—also plausible—way to interpret learning curves. Suppose a subject's strategy is to guess the true hypothesis as to causal strength, but to temper that guess when data are sparse and the causal rating "looks to be" strongly positive or negative, so as to minimize the chances of being mistaken or very badly mistaken. If, for example, h_1 is the most likely hypothesis on the evidence, but it is only minimally more likely than other hypotheses, one could minimize the risk of being (very) wrong by making an estimate closer to h_9. Such a conservative strategy has obvious payoffs where significant costs attend a wrong guess.[11] The existing data, therefore, do not favor the R–W theory over any of several Bayesian interpretations. [12]

[11] Here, at the formal level, it becomes necessary to invoke decision theory. For a treatment of the risk of adopting a false hypothesis, see the work of Neyman and Pearson, e.g., in Neyman (1950). It deserves mention that learning curves represent averages over several subjects; their smoothness can be misleading. Sometimes, individuals show sudden and dramatic increases in learning from one trial to the next, after a number of trials which produce little learning. Though there are mechanistic explanations of this phenomenon (see, e.g., Hilgard and Bower, 1948/1966, pp. 364–375), there are some rather natural rationalistic explanations: the individual has "caught on" to the mechanism involved or to the existence of a correlation; or it has suddenly realized the relevance of some previous learning to the problem which confronts it.

[12] Baker, Berbrier, and Vallee–Tourangeau (1989) show that Shanks' (1985) experiments confirming R–W predictions with human subjects contain possible methodological flaws. Baker et al. considered that subjects might initially average experienced frequencies with prior probability assignments. They confirmed some of Shanks' results with an improved experiment, but did not obtain learning curves, even though they asked subjects to give prior estimates of probabilities. However, the experimental situation and manner in which these priors were elicited provide no reason to expect that subjects took their estimates of priors seriously. It may well be that in sufficiently unfamiliar circumstances, which makes the invention of causal hypotheses and the assignment of priors to them sheer guess work, humans resort (at least initially) to enumerative induction, which does not generate learning curves. The resulting probability estimates derived from relative frequencies in an initial sample of experiences may in turn generate causal hypotheses; and subjects could then "cross over" to Bayesian reasoning. We know of no data suggesting that this does (or does not) occur. For more recent measurements of blocking and conditioned inhibition in human subjects using an experimental design that avoids the flaws in Shanks (1985), see Chapman and Robbins (1990). It is worth remarking that such studies, which are designed to study the manner in which elements of compound cues compete for predictive strength, typically ignore issues which would arise in the mind of a subject who reasoned in a properly scientific manner. Shanks had subjects estimate the relative likelihood that shells and landmines would blow up tanks (in a computer simulation). Gluck and Bower (1988) had subjects estimate the predictive value of various symptoms for the presence of a disease. Chapman and Robbins asked subjects to use the rise in price of fictitious stocks as an indicator of a general rise in stock prices. In the first of these cases, subjects are required to estimate the likelihood that certain causes will be followed by a given effect. In the second, subjects must reason from effects to causes. But in the third, one possible effect of some cause (of stock price fluctuations) is to be used to estimate the likelihood of another possible effect of that cause. These three cases involve different kinds of reasoning and permit different sorts of explanatory hypotheses for what occurs when cues compete. Nor are they probabilistically symmetric: for example, if A and B are effects of a common cause, their co-occurrence will be more probable than the product of their individual a priori probabilities of occurring. If, on the other hand, A and B are individually or jointly sufficient causes of a common effect, the probability of their co-occurrence will not be greater than the product. For a discussion of these asymmetries, see Salmon (1984). They have been used to good effect by Waldmann and Holyoak (1990) in designing experiments to show that people do, at least sometimes, reason in terms of a hypothesized representation of cause-effect relationships, rather than in the manner predicted by mechanical/associationistic models.

Before discussing the plausibility of the Bayesian approach on more general grounds, we shall briefly consider whether there are experiments which would enable us to distinguish the behaviors of R–W organisms from those of Bayesian reasoners. Although clearly no single result could be decisive, there are a number of experiments which are relevant to testing the two models. We summarize here three of these, with the differing predictions made by the R–W model and the Bayesian one.

(1) A^+/X^+, each trained to asymptote; then AX^+ to asymptote. (Then test for strength of conditioning to A and/or to X.)

Since, on the R–W theory, the US supports a maximum associative strength, A and X will each acquire strength at asymptote when presented separately. When they are jointly presented, their conjunction AX acquires strength, and since $V_{AX}=V_A+V_X$, the associative strength of one of A and X must now decrease to half its original value, or less. A Bayesian subject, on the other hand, first learns that A and X are each singly sufficient to produce the US. When trained with AX^+, the subject might entertain the hypothesis that A and X partly counteract one another (thus explaining why the US is no stronger than before); but the new training would only very weakly disconfirm the hypotheses that A and X were singly sufficient for the US, in favor of the alternative that each is only weakly effective (so that together they are very effective). Though an eligible hypothesis, this would be rendered very unlikely by the previous long runs of A^+ and X^+.

(2) Three groups are trained as follows, and the rates of learning under the condition CS^+ are compared:

(a) CS^-; CS^+

(b) CS^+

(c) $CS–CS_1 \ldots CS–CS_n$; CS^+

Group (a) is given preexposure to some appreciable number *n* of occurrences of the CS, unpaired with any US. According to the R–W theory, the absence of a US in this condition means that $\lambda=0$; hence, learning under the subsequent CS^+ condition ought to match that of group (b).[13] But a Bayesian placed under schedule (a) would first learn that the CS did not signal any change in background conditions, and would subsequently have to unlearn that conclusion—hence the learning rate under CS^+ should be better under schedule (b) than under (a). There is, however, another mechanistic theory, due to Mackintosh, which would agree with the Bayesian prediction: on that theory, preexposure to a CS unpaired with any US decreases the amount of attention the subject pays to the CS; and learning rates are dependent upon

[13] It is already known that conditioning to a CS is slower under schedule (a) than under (b). Wagner and Rescorla (1972) modify the R–W theory to accommodate this fact by allowing that the CS^- trials in (a) lower the value of α_{cs}; in effect, they concede that unreinforced CSs may lose salience by ceasing to attract attention.

the degree of attention a CS commands. But the attentional theory holds that attention is a function of the extent to which the CS signals unexpected subsequent events. Thus, if the CS is sequentially paired with *n* novel CSs as in (c), attention ought to be maintained, and learning rates for CS^+ under schedules (b) and (c) ought to match. Not so for a Bayesian organism: it learns on schedule (c) to expect the unexpected when CS occurs; thus, the subsequent *regular* pairing of CS with + ought to be learned more slowly than for a subject placed in condition (b).

We know of no one who has performed experiment (1) above, or compared case 2(c) with 2(b); but, the following test (3) has been performed with rats—though not with a view to comparing the Rescorla–Wagner and Bayesian models.

(3) Train three groups of rats as follows, and then test the degree of conditioning to the stimulus X:

(a) AX^+; A^+
(b) AX^+
(c) AX^+; A^-

We recall that A^+; AX^+, the reverse of (a), is just the blocking experiment. For a Bayesian (who in effect uses Mill's method of agreement here), the order of presentation of A^+ and AX^+ should not matter; in either case, conditioning to X will be lower than for group (b). Similarly, schedule (c) provokes the employment, in effect, of Mill's method of difference; a Bayesian will be more strongly conditioned to X under this schedule than under either (a) or (b). But the Rescorla–Wagner theory predicts equal conditioning to X in all three cases—thus, the reversed order in (a) of A^+ and AX^+ will eliminate blocking. To see this, we must see that the associative strength of X in (a), (b), and (c), once acquired, is not affected in any way by subsequent presentations of A^+ or of A^-; these serve only to change the associative strength of A.

A comparison of schedules (a) and (b) and (c) in (3) has been reported by Kalat and Rozin (1972) for rats. Kamin (1969), among others, also working with rats, compared schedule (a) to (b). In both cases, the results support the Rescorla–Wagner theory over the more cognitive Bayesian theory.

Non-Bayesian rats aside, there are a number of further issues which a Bayesian theory must address. Thus, for example, it seems plausible to handle overshadowing phenomena by assigning higher priors to hypotheses which assign the causal potency of a compound CS primarily to the stronger component(s). There is also the question of what determines the competing hypotheses a subject entertains when confronting a given set of data. This is a difficult question; here, we offer just two principles which we believe have psychological plausibility.

1. Confronted by a US which is statistically paired with one or more simple or complex CSs, a subject will entertain hypotheses which attribute causal power to the constituent CSs, singly or in the various combinations found within complex CSs; unper-

> ceived causes are not hypothesized except when USs unpaired with any observed CS occur.
>
> II. The possibility that each of two or more elements of a complex CS is alone causally sufficient to produce the US is discounted unless there is independent evidence of this. Discounting, as this is called, is a kind of simplicity postulate: crudely, it enjoins multiplication of causes (or causal strength) beyond necessity. There is some empirical evidence for this bias against causal overdetermination (see Sedlak and Kurtz, 1981).

We have provided here only a very partial sketch of a Bayesian approach to causal thinking. Rather than attempting to develop it further, we believe it will be most useful to conclude by addressing some very general issues which bear upon the plausibility of mechanistic versus cognitive explanations of the capacity to apprehend causal connections.

It is obvious that *Hermissenda* do not apply Bayes' theorem or even calculate relative frequencies. But, to be sure, neither do human beings—except in the most rarified of contexts. Not even experimental psychologists and others trained in the relevant statistical techniques will engage in anything so complex as the calculations just outlined, except when they need to publish the results of their studies. Thus, initially, it appears that a Bayesian theory will have no plausibility as a realistic account of the acquisition of causal knowledge in ordinary circumstances—even for humans.

Does it follow that some mechanistic model (whether it be the R–W theory or some alternative) must be correct? Of course not. What we ought to do is to use models such as the Bayesian one as heuristic devices in the construction of simpler, qualitative inductive strategies which can more reasonably be imputed to humans and perhaps to some lower animals. It is easy enough, for example, for mathematically naive human subjects to appreciate the qualitative features of Bayes' theorem—e.g., that results, predicted by the hypothesis as very likely, tend to confirm it, unless those results were very likely to occur in any case. A variety of other inductive rules can be formulated, more or less powerful, which may capture the reasoning strategies, where reasoning is present, which lead to causal beliefs.[14]

It is fair to say that much remains to be discovered about the processes by means of which causal relations are apprehended and about the conditions under which these processes are reliable. The role of tactual perception of forces in causal reasoning, the development of causal reasoning in children, and the dynamics of belief formation and revision under the influence of increasing experience: all these and much else are as yet poorly understood.

[14] Some of these may be correct, if rough, and others fallacious, but effective in limited contexts. See, in this connection, Shaklee (1983). Einhorn and Hogarth (1986) review a number of theoretical perspectives on rational causal attributions and review relevant empirical studies. Cheng and Novik (1990) make use of what they regard to be a computationally realistic form of Mill's methods to predict the causal attributions of human subjects.

Indeed, it is not easy to discriminate experimentally between the action of mechanical processes and intellectual ones in rendering organisms sensitive to the causal features of their environment, for, as we have seen, rather simple mechanisms are remarkably able to mimic the results of intelligent thought.

Measured against such norms as Mill's principles or Bayesian reasoning, lower animals in many situations perform remarkably well. (For a recent comparison of human versus animal causal learning, and discussion of mechanical versus ratiocinative mechanisms, see Wasserman, 1990. See also Gluck and Bower, 1988, for a discussion of the ability of adaptive networks to achieve computer modelling of learning phenomena.) However, rats, at least, perform poorly by these standards in other situations, as we have just seen. It would be most interesting to learn how well higher animals—e.g., monkeys and mathematically naive humans—perform such tasks. According to Shaklee (1983), they often do none too well. If we judge the reliability of an organism's discrimination of causal relations by how well its performance matches that of a Bayesian or even that of a reasoner using Mill's methods, we may find that humans, in their everyday affairs, fall well short of the ideal. Nevertheless, sophisticated calculations are reliable and effective only to the degree that they can be executed without error and quickly enough to serve our ends. Perhaps Hume was not far wrong when he suggested that nature would have been foolhardy had she left us to rely on so weak and fallible a faculty as that of reason for the purpose of acquiring causal beliefs.

References

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

Anscombe, G.E.M. (1971). *Causality and determination*. London: Cambridge University Press.

Baker, A., Berbrier, M., and Vallee–Tourangeau, F. (1989). Judgments of a 2x2 contingency table: Sequential processing and the learning curve. *Quarterly Journal of Experimental Psychology*, *41B*, 65–97.

Berzonsky, M. (1971). The role of familiarity in children's explanations of physical causality. *Child Development*, *42*, 705–412.

Bullock, M., and Gelman, R. (1979). Preschool children's assumptions about cause and effect: Temporal ordering. *Child Development*, *50*, 89–96.

Bullock, M., Gelman, R., and Baillargeon, R. (1982). The development of causal reasoning. In W. Friedman (Ed.), *The developmental psychology of time* (pp. 209–254). New York: Academic Press.

Burkes, A. (1977). *Chance, cause, and reason: An inquiry into the nature of scientific evidence*. Chicago: The University of Chicago Press.

Carnap, R. (1952). *The continuum of inductive methods*. Chicago: The University of Chicago Press.

Chapman, G., and Robbins, S. (1990). Cue interaction in human contingency judgment. *Memory and Cognition*, *18*, 537–545.

Chatlosh, D.L., Neunaber, D.J., and Wasserman, E.A. (1985). Response-outcome contingency: Behavioral and judgmental effects of appetitive and aversive outcomes with college students. *Learning and Motivation*, *16*, 1–34.

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

Dickinson, A., Hall, G., and Mackintosh, N.J. (1976). Surprise and the attenuation of blocking. *Journal of Experimental Psychology: Animal Behavior Processes, 2*, 313–322.

Dickinson, A., and Mackintosh, N.J. (1979). Reinforcer specificity in the enhancement of conditioning by posttrial surprise. *Journal of Experimental Psychology: Animal Behavior Processes, 5*, 162–177.

Domjan, M. (1983). Biological constraints on instrumental and classical conditioning: Implications for general process theory. *The Psychology of Learning and Motivation, 17*, 215–277.

Ducasse, C.J. (1951). *Nature, mind, and death*. LaSalle, Illinois: Open Court.

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

Elek, S. (1990). *Response to and report of response-outcome contingencies: A developmental study*. Unpublished doctoral dissertation. University of Iowa, Iowa City, Iowa.

Fales, E. (1990). *Causation and universals*. New York: Routledge.

Farley, J. (1987a). Contingency learning and causal detection in *Hermissenda:* I. Behavior. *Behavioral Neuroscience, 101*, 13–27.

Farley, J. (1987b). Contingency learning and causal detection in *Hermissenda:* II. Cellular mechanisms. *Behavioral Neuroscience, 101*, 28–56.

Gemelli, A., and Capellina, A. (1958). The influence of the subject's attitude in perception. *Acta Psychologica, 14*, 12–23.

Gibbon, J. (1981). The contingency problem in autoshaping. In C.M. Locurto, H.S. Terrace, and J. Gibbon (Eds.), *Autoshaping and conditioning theory* (pp. 285–308). New York: Academic Press.

Gibbon, J., and Balsam, P. (1981). Spreading association in time. In C.M. Locurto, H.S. Terrace, and J. Gibbon (Eds.), *Autoshaping and conditioning theory* (pp. 219–253). New York: Academic Press.

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

Golnikoff, R.M., Harding, C.G., Carlson, V., and Sexton, M.E. (1984). The infant's perception of causal events: The distinction between animate and inanimate objects. In L.P. Lipsitt and C. Rovee-Collier (Eds.), *Advances in infancy research* (pp. 145–151). Norwood, New Jersey: ABLEX Publishing Corp.

Gormezano, I., and Kehoe, E.J. (1981). Classical conditioning and the law of contiguity. In Harzem and M.D. Zeiler (Eds.), *Predictability, correlation, and contiguity* (pp. 1–45). New York: John Wiley & Sons.

Granger, R.H., and Schlimmer, J.C. (1986). The computation of contingency in classical conditioning. *The Psychology of Learning and Motivation, 20*, 137–192.

Grüber, C., Fink, D., and Damm, V. (1957). Effects of experience on perception of causality. *Journal of Experimental Psychology, 53*, 89–93.

Harré, R., and Madden, E. (1975). *Causal powers*. Totowa, New Jersey: Rowman and Littlefield.

Hilgard, E.R., and Bower, G.H. (1966). *Theories of learning* (third edition) New York: Appleton–Century–Crofts. (Originally published 1948)

Hume, D. (1888). *A treatise of human nature*. [L.A. Selby-Bigge, Ed.]. London: Oxford University Press. (Originally published 1739)

Hume, D. (1955). *An inquiry concerning human understanding*. [C. Hendel Ed.]. Indianapolis: Bobbs-Merrill. (Originally published 1758)

Jenkins, H.M., Barnes, R.A., and Barrera, F.J. (1981). Why autoshaping depends on trial spacing. In. C.M. Locurto, H.S. Terrace, and J. Gibbon (Eds.), *Autoshaping and conditioning theory* (pp. 255–284). New York: Academic Press.

Kalat, J.W., and Rozin, P. (1972). You can lead a rat to poison but you can't make him think. In M. Seligman and J. Hager (Eds.), *Biological boundaries of learning* (pp. 115–122). New York: Appleton–Century–Crofts.

Kamin, L.J. (1969). Predictability, surprise, attention, and conditioning. In B. Campbell and R. Church (Eds.), *Punishment and aversive behavior* (pp. 279–296). New York: Appleton–Century–Crofts.

Kimble, G.A. (1961). *Hilgard and Marquis' conditioning and learning* (second edition). New York: Appleton–Century–Crofts.

Leslie, A.M. (1982). The perception of causality in infants. *Perception, 11*, 173–186.

Mackintosh, N.J. (1971). An analysis of overshadowing and blocking. *Quarterly Journal of Experimental Psychology, 23*, 118–125.

Mackintosh, N.J. (1974). *The psychology of animal learning*. London: Academic Press.

Maine de Biran, P. (1942). *Oevres choisies de Main de Biran*. Paris: Aubier.

Mendelson, R., and Shultz, T.R. (1976). Covariation and temporal contiguity as principles of causal inference in young children. *Journal of Experimental Child Psychology, 22*, 408–412.

Michotte, A. (1963). *The perception of causality*. [T. Miles and E. Miles, Trans.]. London: Methuen. (Originally published 1946)

Mills, J.S. (1911). *A system of logic* (eighth edition). London: Longmans, Green and Co. (Originally published 1843)

Neyman, J. (1950). *First course in probability and statistics*. New York: Henry Holt.

Olum, V. (1956). Developmental differences in the perception of causality. *American Journal of Psychology, 59*, 417–425.

Olum, V. (1958). Developmental differences in the perception of causality under conditions of specific instructions. *Vita Humana, 1*, 191–203.

Piaget, J. (1966). *The child's conception of causality* [M. Gabain, Trans.]. London: Keegan Paul. (Originally published 1930)

Piaget, J. (1974). *Understanding causality* [D. Miles and M. Miles, Trans.]. New York: W. W. Norton. (Originally published 1971)

Powesland, F. (1959). The effect of practice upon perception of causality. *Canadian Journal of Psychology, 13*, 155–168.

Rescorla, R.A., and Wagner, A.R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. Black and W. Prokasy (Eds.), *Classical conditioning II: Current research and theory* (pp. 64–69). New York: Appleton–Century–Crofts.

Russell, B. (1917). On the notion of cause. In B. Russell, *Mysticism and logic*. Totowa, New Jersey: Barnes and Noble Books.

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

Sedlak, A.J., and Kurtz, S.T. (1981). A review of children's use of causal inference principles. *Child Development, 52*, 759–784.

Shaklee, H. (1983). Human covariation judgment: Accuracy and strategy. *Learning and Motivation, 14*, 433–448.

Shanks, D.R. (1985). Continuous monitoring of human contingency judgments across trials. *Memory and Cognition, 13*, 158–167.

Shanks, D.R. (1987). Acquisition functions in contingency judgment. *Learning and Motivation, 18*, 147–166.

Shultz, T.R. (1980). Development of the concept of intention. In W. Collins (Ed.), *The Minnesota symposium on child development* (Vol. 13, pp. 131–164). Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Shultz, T.R. (1982). Rules of causal attribution. *Monographs for Research in Child Development, 47*, 1–51.

Siegler, R.S. (1975). Defining the locus of developmental differences in children's causal reasoning. *Journal of Experimental Child Psychology, 20*, 512–525.

Siegler, R.S., and Liebert, R.M. (1974). Effects of contiguity, regularity, and age on children's causal inferences. *Developmental Psychology, 10*, 574–579.

Skinner, B.F. (1948). "Superstition" in the pigeon. *Journal of Experimental Psychology, 38*, 168–172.

Uzgiris, I.C. (1984). Development in causal understanding. In L. Lipsitt and C. Rovee-Collier (Eds.), *Advances in infancy research* (pp. 130–135). Norwood, New Jersey: ABLEX Publishing Corp.

Wagner, A.R., and Rescorla, R.A. (1972). Inhibition in Pavlovian conditioning: Application of a theory. In R. Boakes and M. Halliday (Eds.), *Inhibition and learning* (pp. 301–306). New York: Academic Press.

Waldmann, M.R., and Holyoak, K.J. (1990). *Can causal induction be reduced to associative learning?* Technical Report UCLA-SCRP-90-6.

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

Wasserman, E.A., and Neunaber, D.J. (1986). College students' responding to and rating of contingency relations: The role of temporal contiguity. *Journal of the Experimental Analysis of Behavior, 46*, 15–35.

Wasserman, E.A., Chatlosh, D.L., and 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.

Watson, J.S., and Ramey, C.T. (1972). Reactions to response-contingent stimulation in early infancy. *Merrill-Palmer Quarterly, 18*, 219–227.

Watson, J.S. (1984). Bases of causal inference in infancy: Time, space, and sensory relations. In P. Lipsitt and C. Rovee-Collier (Eds.), *Advances in infancy research* (pp. 152–160). Norwood, New Jersey: ABLEX Publishing Corp.

Yéla, M. (1952). Phenomenal causation at a distance. *Quarterly Journal of Experimental Psychology, 4*, 139–154.