Proceedings of the Cognitive Science Society Conference, 1986

Conditioning and categorization
Some common effects of informational variables in animal and human learning

Mark A. Gluck & Gordon H. Bower
Stanford University

To what extent do the processes of human learning emerge from complex configurations and elaborations of the "elementary" learning processes observed in animals? Research in the two areas of human and infra-human learning share a long history which focussed on elementary associative learning (Ebbinghaus, 1885; Pavlov, 1927). About twenty years ago, however, animal and human learning research became divorced from each other. Animal research continued to be primarily concerned with elementary associative processes (Mackintosh, 1983; Mackintosh & Honig, 1969; Rescorla & Holland, 1982); while human learning (or "memory") tended to be characterized in terms of information-processing and rule-based, symbol-manipulation, an approach borrowed from artificial intelligence. Few current theories of learning attempt to bridge the gap between human and infra-human learning (some exceptions include Alloy & Tabachnick, 1984; Estes, 1985; Dickinson & Shanks, 1985; Medin, 1984; Holland, Holyoak, Nisbett, & Thagard, in press). Recently, however, interest in relating human cognition to configurations of elementary associative connections has revived. Among theorists using parallel-distributed processing models, the works of McClelland, Rumelhart, Hinton, Sejnowski, and James Anderson are notable for demonstrating the computational power and psychological verisimilitude of these "connectionist" networks (see e.g., Hinton & Anderson, 1981; McClelland & Rumelhart, 1981; Ackley, Hinton, & Sejnowski, 1985, Rumelhart & McClelland, 1986).

Given the voluminous studies of learning in animals alongside current attempts to model cognition with elementary associative processes, it would seem particularly timely to search for and exploit any correspondences which might exist between animal and human associative learning. This was our goal in these experiments.

Informational Variables in Classical Conditioning

A simple but powerful theory describing animals' learning in classical Pavlovian conditioning was presented by Rescorla and Wagner in the early 1970's (Rescorla & Wagner, 1972; Wagner & Rescorla, 1972). In classical conditioning, a previously neutral stimulus, the *conditioned stimulus* (CS), such as a bell, comes to be associated with a biologically significant stimulus, the *unconditioned stimulus* (US), such as food or an electric shock. Early learning theories assumed that the simple temporal contiguity or joint occurrence of a CS and US was *sufficient* for associative learning (e.g. Hull, 1943; Spence, 1956). Later experiments made clear, however, that simple contiguity was not sufficient. The ability of a CS to become conditioned to a US depended on its imparting reliable and non-redundant information about the occurrence of the US (Kamin, 1969; Rescorla, 1968; Wagner, 1969).

To illustrate, suppose that a light, the CS, has already been conditioned to predict a shock, the US. If a compound stimulus consisting of a light and a tone is then paired with the shock, learning of the *tone→shock* association hardly occurs at all compared to control subjects who received no pretraining to the light (Kamin, 1969). This result, similar to Pavlov's work on

the overshadowing of one cue by another, is called "blocking" because prior training of the *light→shock* association blocks later learning of the *tone→shock* association during the second, *(light + tone)→shock* stage of training.

The Rescorla-Wagner Model

The blocking effect suggested that the effectiveness of a US for producing associative learning depends on the relationship between the CS and the *expected outcome* (Rescorla, 1968; Wagner, 1969; Kamin, 1969). Rescorla and Wagner provided a precise formulation of this proposal (Rescorla & Wagner, 1972; Wagner & Rescorla, 1972). Their formulation assumes that the association which accrues between a stimulus and its outcome on a trial is proportional to the degree to which the outcome is unexpected (or unpredicted) given *all* the stimulus elements that are present on that trial. We let V_i denote the strength of association between stimulus element CS_i and the *US*. If CS_i is followed by a reinforcing unconditioned stimulus, *US*, then the change in the association strength between CS_i and the *US*, ΔV_i, can be described by Equation (1):

$$\Delta V_i = \alpha_i\beta_1(\lambda_1 - \sum_{k\epsilon S} V_k), \tag{1}$$

where α_i reflects the intensity or salience of CS_i, β_1 reflects the rate of learning on trials with US presentations, λ_1 is the maximum possible level of association strength conditionable with that US intensity, and $\sum_{k\epsilon S} V_k$ is the sum of the associative strengths between all the CS stimulus elements occuring on that trial and the US. If CS_i is presented on a trial without the US, then the association between CS_i and the US decreases analogously, viz.,

$$\Delta V_i = \alpha_i\beta_2(\lambda_2 - \sum_{k\epsilon S} V_k), \tag{2}$$

where λ_2 is the level of associative strength supported by non-presentation of the US (usually taken to be zero), and β_2 reflects the rate of change of the association due to nonreinforcement. Generally β_1 is assumed to be larger than β_2, but this is not critical for most predictions (see Rescorla & Wagner, 1972).

The Rescorla-Wagner model is the most widely accepted description of associative changes during classical conditioning. The wealth of confirmed implications arising from this deceptively simple model has been substantial. This model accounts for the blocking effect as follows: When in Phase 1, CS_1 has been initially conditioned to the US, V_1 approaches λ_1. If the associative strength of the novel stimulus, V_2 is assumed to be zero, then the *compound stimulus strength*, $V_1 + V_2 = \lambda_1$. By Equation 1, the incremental learning accruing to the novel stimulus, ΔV_2, when the compound is paired with the US is thus predicted to be zero--as observed.

Learning in Associative Networks

A learning rule used in many of the "connectionist" network models of cognition is the *delta rule*, a variant of the perceptron convergence procedure (Rosenblatt, 1961) first proposed as a learning mechanism for adaptive networks by Widrow and Hoff (1960). Such networks connect a set of input nodes to some output nodes with "connection weights" w_{ij} from node i to node j. Given a training trial relating an input vector to an output vector, the weights are changed according to (3):

$$\Delta w_{ij} = \beta(z_j - \sum_{k=1}^{n} w_{kj} a_k) a_i, \tag{3}$$

where i is an input node, j is an output node, a_i is the activation on input node i, the summation is over all the input nodes to node j, and z_j is a special "teaching" input signal to output node j indicating what the activation of that node should be to get the correct response. The delta rule provides an iterative solution to a set a linear equations which will converge on discriminating weights if they exist. Otherwise, the algorithm will converge on weights which minimize the "least-squares" error between the resulting and desired output patterns (Kohonen, 1977).

Recently, Rumelhart, Hinton, and Williams (1986) have generalized the delta rule so it may be applied to perform learning in a multi-layered net of feed-forward elements with some "hidden" units between the input and output layers. They show how the delta rule, combined with back-propagation of weight adjustments, can learn many difficult discriminations such as parity, exclusive-or, and symmetry relationships.

As Sutton and Barto (1981) noted, the delta rule is essentially identical to the Rescorla-Wagner equations (with $\beta_1 = \beta_2$). If, in Equation 3, we let $V_i = w_{ij}$, set the training signal in the delta rule, z_j, equal to λ_1 when the US is present and to zero otherwise, and let a_i=1 when CS_i is present and 0 otherwise, then the delta rule reduces to Equations 1 and 2 of the Rescorla-Wagner model. Curiously, associative network theorists have adopted the delta rule because of its computational power, convergence properties, and generalizability to multi-layered networks. Nonetheless, associative networks which implement the delta rule can be viewed as a framework for modeling the emergent properties of complex configurations of elementary associative processes observed in animals. However, few studies have asked whether the delta rule is an appropriate characterization of the algorithm underlying human associative learning.

Some earlier investigators have noted the need for bridging experiments. Rudy (1974) noted a parallel between human paired-associate learning and animal associative learning and pointed to a form of blocking in human learning. Specifically, when redundantly relevant cues are compounded with stimuli that are already sufficient to cue the associated response, the added cues are unlikely to become associated with the response (Trabasso & Bower, 1968). Dickinson and Shanks (1985) demonstrated some conditioning phenomena in human learning: They showed that human judgments of event correlation were influenced by the conditional status of other events that are present, in a manner reminiscent of blocking or overshadowing phenomena in animal conditioning. Schank (1982) has recently postulated a similar "expectation failure" as

the driving force behind learning; EPAM used a similar rule long ago (Feigenbaum, 1959; Feigenbaum & Simon, 1961).

Experiment 1

Because category learning is a currently active area in cognitive research, we decided to test out the delta rule as it applied to subjects' learning to classify stimulus patterns into categories. In our experiment, university students served as hypothetical medical diagnosticians. They saw a series of 250 "patients," each described by the presence or absence of each of four symptoms. The student diagnostician classified each patient as having one or the other of two fictitious diseases, received feedback about that patient's correct diagnosis. Over training, subjects learned which symptoms are more or less diagnostic of which diseases.

Figure 1 illustrates a simple associative network to represent this category learning. Each of the four symptoms is represented by an input node at the left, and the two disease categories by nodes at the right. The connections from symptom i to category j has weight, w_{ij}, reflecting the strength of evidence that presence of symptom i provides towards disease j. The w_{ij} will be adjusted trial by trial according to the delta rule.

The pattern of features presented on a trial causes a pattern of activation of the features. If the presence or absence of each feature is represented by activations of 1 and 0, respectively, the activation at a given category node will equal the sum of the weights from presented features to that category node. This reflects the model's expectation for that category given the symptom pattern. Once activation values are computed for the category nodes, the next step is for the model to select a response. Several measures of associative strength are possible. One we have used is to ask subjects to judge directly the probability that a given patient has one disease or the other. We will suppose that the greater the difference in net strength of evidence for category 1 vs. 2, the higher will be subject's estimate that the patient has disease 1 rather than disease 2. A second measure asks subjects to choose disease 1 or 2 for a particular patient. For this case, we use the ratio response rule of Luce (1963) which says that the probability of choosing Category 1 is the ratio $\frac{V_1}{(V_1 + V_2)}$. Qualitative aspects of the predictions do not depend on the details of the response rule.

The "training signal" provided to each category node (Figure 1) is the experimenter's feedback (after the subject's response) regarding the correct response. We assume that if category j is the correct classification, then z_j will be set equal to one on that trial; if an alternative category is correct on a given trial, then z_j will be 0 for that trial. We assume a single learning rate parameter, β, for adjusting the weights. However, if a fixed set of training patterns is presented many times in random order to the learning model, the convergence properties of the delta rule lead to *parameter free* predictions about the expected asymptotic levels of the w_{ij}'s, feature-to-category associations (Rescorla & Wagner, 1972; Stone, 1986).

We will compare the predictions of the delta rule in our category learning task with the predictions of three competing models of category learning (Estes, 1986): 1) *exemplar* models which presume that the learner stores all the exemplars of each category, and then classifies a new instance according to its similarity to the stored exemplars of each category (e.g. Medin &

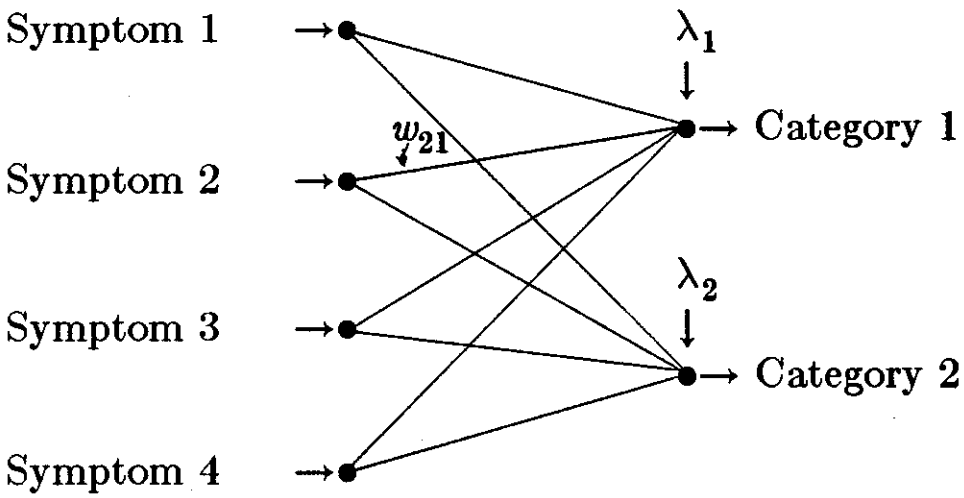

Figure 1. A simple "connectionist" network which learns to diagnose patterns of up to four symptoms as having one of two diseases using the Rescorla-Wagner/delta rule.

Schaffer, 1978; Nosofsky, 1984), 2) *feature-frequency* which presume that the learner stores relative frequencies of occurrence of cues within the categories, and then classifies an instance according to the relative likelihood of its particular pattern of features arising from each of the categories (Reed, 1972; Franks & Bransford, 1971), and 3) *prototype* models which presume the learner abstracts the central tendency (modal description) of each category and then classifies instances according to their similarity to this central prototype (Fried & Holyoak, 1984).

Applying models to our task where subjects estimate the probability of each category given each feature, the models make one of two predictions. Exemplar models and feature-frequency models predict that subjects' estimates will simply reflect the observed conditional feature-to-category probabilities of the training sequence, a form of "probability matching." On the other hand, prototype models and feature-frequency models which ignore variations in category base-rate frequencies would predict that subjects' estimates of the probability of the category given the feature will reflect simply the relative likelihood of the feature given the alternate categories, viz.,

$$\frac{P(f|c_1)}{P(f|c_1)+P(f|c_2)}.$$

In our experiment, we arranged to have the ordinal relationships among the conditional probabilities for different cues differ from the ordinal relationships among the expected asymptotic association strengths predicted by the Rescorla-Wagner/delta rule. This was achieved by unbalancing the relative frequency of the two diseases, making the common disease far more likely than the rare disease. The question was whether people's probability estimates would be more closely predicted by the Rescorla-Wagner/delta rule than by the alternative models.

Procedure

Nineteen subjects were trained to classify medical charts of hypothetical patients into one of two mutually exclusive disease categories. Disease names were fictitious but we will refer to them as the rare (R) disease and the common (C) disease. Among the training exemplars, patients with the common disease were three times as frequent as patients with the rare disease. A patient chart consisted of one to four symptoms drawn from a set of four possible symptoms: bloody nose, stomach cramps, puffy eyes, and discolored gums. In the training phase subjects were shown a set of symptoms corresponding to a patient, asked to make a diagnosis, and then given feedback as to the correct diagnosis. Figure 2a shows the probability of each of the four symptoms occurring in patients suffering from each of the two diseases. The lower numbered symptoms were more typical for the rare disease while the higher numbered symptoms were more typical of the common disease. All symptoms, however, occurred in some patients with both diseases. Symptoms 1, 2, 3, and 4 were assigned actual symptom names randomly for each subject. Each subject received a novel set of training patients which were generated during the experiment according to a probabilistic procedure. First, each patient was randomly designated as suffering from either the rare disease (with probability .25) or the common disease (with probability .75). Second, given his disease, a patient's symptom chart was generated by choosing symptoms according to the *independent* probabilities shown in Figure 2a. Thus, if the patient suffered from the rare disease, then with probability .6, the chart would include symptom 1; with probability .4, symptom 2; with probability .3, symptom 3; and with probability .2, symptom 4 (and analogously, but inversely, for patients suffering from the common disease). From one to four symptoms were presented on a single chart (patients with no symptoms were eliminated

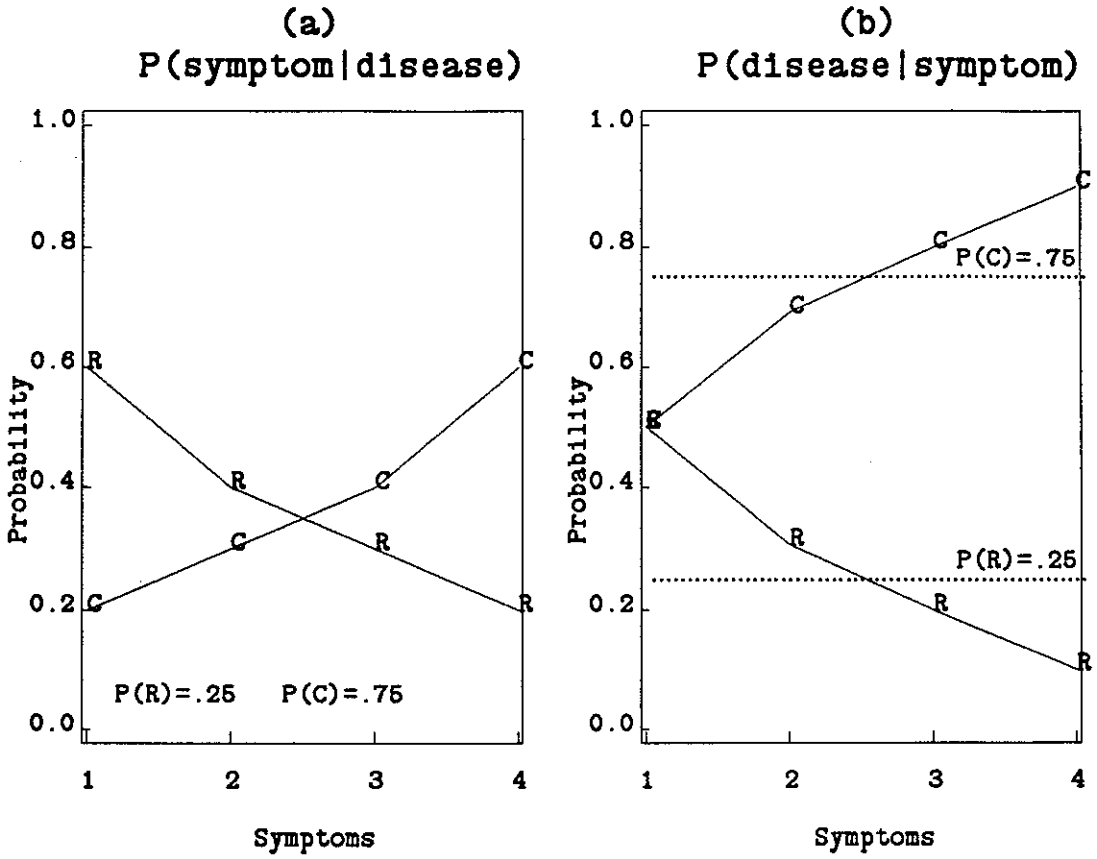

Figure 2. Experiment 1 design: (a) The probabilities of each of the four symptoms occurring in patients suffering from each of the two diseases. The lower numbered symptoms were more typical for the rare disease while the higher numbered symptoms were more typical of the common disease. (b) The conditional probabilities of each of the two diseases given the presence of each of the symptoms computed from (a) using Bayes Theorem.

from the training sequence). For the subjects, the diseases were identified by fictitious names which were counterbalanced across subjects in being assigned to the rare or common disease. Subjects were instructed that there was no simple rule for making the diagnosis and that the order of presentation of the symptoms within a patient's chart was irrelevant.

Using the base rates of $P(R) = .25$ and $P(C) = .75$ and the probabilities in Figure 2a, Bayes Theorem provides the conditional probability of the two diseases given the four symptoms considered seperately (see Figure 2b). For any single symptom the normative probability of the rare disease was always less than or equal to the probability of the more common disease.

Following 250 training trials of predicting diseases and receiving feedback, subjects were finally asked to estimate directly the probability that a patient exhibiting a particular symptom was suffering from one or the other disease. They gave a numerical estimates of $P(R|s_i)$ and $P(C|s_i)$ on a 0 to 100 scale for each of the four symptoms. These estimates are the data of primary interest in this report.

Results and Predictions

Because the conditional probabilities of the two diseases sum to 1 for any particular symptom, we will combine these conditional probabilities into a single *probability difference measure*, $P(R|s_i) - P(C|s_i)$, for each of the four symptoms. This measure, shown in Figure 3, reflects both the actual (normative) probabilities in the training patterns as well the probability matching behavior predicted by exemplar-storage and feature-frequency models. But, the Rescorla-Wagner/delta rule predicts that following training, subjects' estimates of the probability differences will follow a different pattern, reflecting the underlying strengths of the feature-to-category associative connections. These asymptotic connection weights can be calculated by deriving equations for the expected trial-by-trial weight change in each of the feature-to-category connections, setting these expected changes to zero, and solving the resulting four simultaneous equations in four variables for each of the two categories. The resulting asymptotic association strengths to the rare disease are .45, .18, .06, and -.09 for symptoms 1 through 4, respectively, and for the associations to the common disease, .02, .22, .37, and .68. The differences between these asymptotic strengths are plotted in Figure 3b; this is the theoretical index to be compared to the observed probability difference measures.

The most striking difference between the normative probability measures in Figure 3a and the predicted associative weights in Figure 3b is evident in symptom 1, s_1: This symptom was paired equally often with the rare disease, R, as with the common disease, C, and hence the difference between the conditional probabilities of R versus C to this cue is zero. However, the delta-rule predicts that the the $s_1 \rightarrow R$ association will be considerably stronger than the $s_1 \rightarrow C$ association.

This prediction of the delta rule is understandable when one appreciates the *competitive* nature of the learning algorithm. The overall magnitude of the *symptom→disease* weight reflects the degree to which a symptom has been an informative and reliable predictor of a disease, *relative* to the predictive value of other co-present symptoms for that same disease. Although symptom 1 has the same predictive value for the two diseases, relative to the predictive value of the other symptoms for the common disease, it is not a very informative predictor. However, for the rare disease symptom 1 is a relatively better predictor than the other symptoms. It is this