Extending the Domain of a Feature-based Model of Property Induction

Steven Sloman[1]
Department. of Psychology
University of Michigan
sloman@psych.stanford.edu

Edward Wisniewski
Psychology Department
Northwestern University
Evanston, IL 60208
edw@nwu.edu

Abstract

A connectionist model of argument strength, which applies to arguments involving natural categories and unfamiliar predicates, was proposed by Sloman (1991). The model applies to arguments such as *robins have sesamoid bones, therefore hawks have sesamoid bones.* The model is based on the hypothesis that argument strength is related to the proportion of the conclusion category's features that are shared by the premise categories. The model assumes a two-stage process in which premises are first encoded by connecting the features of premise categories to the predicate. Conclusions are then tested by examining the degree of activation of the predicate upon presentation of the features of the conclusion category. The current work extends the domain of the model to arguments with familiar predicates which are nonexplainable in the sense that the relation between the category and predicate of each statement is difficult to explain. We report an experiment which demonstrates that both of the phenomena observed with single-premise specific arguments involving unfamiliar predicates are also observed using nonexplainable predicates. We also show that the feature-based model can fit quantitatively subjects' judgments of the strength of arguments with familiar but nonexplainable predicates.

Introduction

One of the most striking capacities of the human mind is the ease with which it can generate new beliefs from old ones. One form of this capacity is property-induction: The ability to express degrees of belief that one category of things exhibits some property given that other categories do. This ability can be expressed as a judgment of the strength of an argument in which the premises specify the relevant old beliefs and the conclusion specifies the newly hypothesized category-property relation. An example of such an argument is

i. Robins secrete uric acid crystals.
Penguins secrete uric acid crystals.
Therefore, Hawks secrete uric acid crystals.

How do people transmit belief from the premises to the conclusion of such an argument and what kind of systematicities in human judgment can we expect as a result of this process?

As an alternative to a model proposed by Osherson et al. (1990), Sloman (1991) proposed a simple connectionist network to model the subjective strength of a restricted class of arguments. Each argument consisted of a set of propositions, with all but one taken as statements of fact (premises). Subjects judged the validity of the remaining proposition (the conclusion) in light of the premises. Each proposition consisted of a one-place predicate (e.g., "secretes uric acid crystals") and a natural-kind object-category (e.g., "robins") to which it applied. Within an argument, all propositions shared a single predicate; only the category differed. The task was further constrained by allowing only predicates that were unfamiliar to subjects (such as "secretes uric acid crystals"). Unfamiliar predicates were used because they severely limit subjects' ability to reason about them. This allows theorists to focus on the transmission of belief amongst the categories of an argument, ignoring the role of the predicate. As described in Sloman (1991), the model was able to account for a host of qualitative phenomena involving arguments with unfamiliar predicates and showed good quantitative fits to subjects' ratings of argument strength.

The current work aims to extend the domain of the model to a class of arguments involving familiar

[1]Steven Sloman is now at the Department of Cognitive and Linguistic Sciences, Box 1978, Brown University, Providence, RI 02912

predicates. Limiting ourselves to single-premise arguments that are *specific* (a superordinate that properly includes one category also includes the other), we show that when subjects cannot explain the relation between the categories and the familiar predicate of an argument, they behave in the same way as they do with unfamiliar predicates. Before describing our evidence for this extension of the model's domain, we briefly identify the class of phenomena that the model was designed to account for, and describe the model itself.

Argument Strength Phenomena

Psychologists have identified about a dozen phenomena or general tendencies concerning the subjective strength of arguments involving unfamiliar predicates (cf. Osherson et al., 1990; Rips, 1975; Sloman, 1991). One example is the diversity phenomenon: People prefer arguments whose premises are less similar. To illustrate, people tend to believe that argument i. above is stronger than an argument with more similar premises like "Robins have X, Sparrows have X, therefore Hawks have X." (Because all predicates are unfamiliar, they can be referred to generically as predicate X.) In the course of describing our feature-based model and its extension to familiar predicates below, we outline four other phenomena: feature exclusion, nonmonotonicity, similarity, and asymmetry.

Feature Coverage

The model is based on the hypothesis that the strength of an argument is directly related to the proportion of the conclusion category's features or attributes that it shares with the premise categories -- the extent to which the features of the premise categories *cover* those of the conclusion category. The key assumptions are that all categories can be represented as a list of features, and that these features can be obtained from subjects. Roughly, an argument is strong to the extent that the features of the conclusion category are spanned by the features of the premise categories.

Feature-Based Induction

By representing categories as feature sets, we are able to distribute the representation of a category over a set of variables or units, where each unit represents a particular feature. One advantage of such a representational scheme is that any learning involving a feature of one category will automatically generalize to other categories sharing that feature. To model the transmission of belief from one category to another, we take advantage of this automatic generalization property. Our model consists of a network of n input units, which are used to represent categories consisting of n features each, and a single output unit, which is used to represent the unfamiliar predicate X. In brief, the model posits that premise categories are first encoded as a vector of weights by connecting the units representing the features of the premise categories to the predicate unit. The network is then presented with the conclusion category. The process is illustrated in Figure 1. The value of the output unit upon presentation of the conclusion category is the model of argument strength. Strength is proportional to the extent to which the features of the conclusion category have been connected to the predicate unit by virtue of the encoding of the premises.

We now describe the two stages of the model in more detail. First, premises are encoded by connecting the features of their categories to the

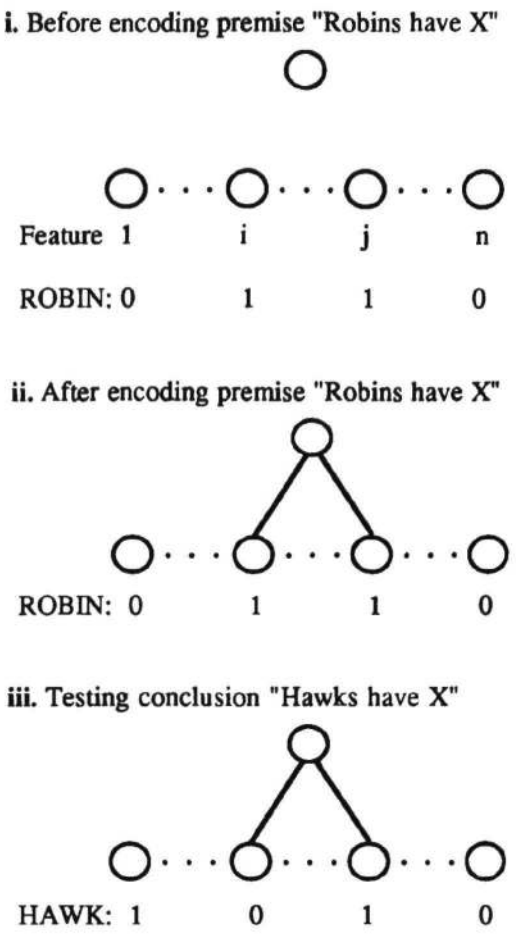

Figure 1. Illustration of the feature-based model for the argument Robins have X, therefore Hawks have X.

predicate unit. The connection at time t from feature i to the predicate unit ($w_{i,t}$) is updated for each premise using the following delta rule:

$$w_{i,t+1} = w_{i,t} + [1 - w_{i,t}][1 - a_X(P)]f_i(P),$$

in which $a_X(P)$ is the activation of the predicate unit upon presentation of premise P and $f_i(P)$ is the value of feature i of the premise's category. The coefficient $[1 - w_{i,t}]$ is used in place of the usual learning-rate parameter to keep each weight between 0 and 1.

In the second stage, the conclusion is tested by presenting its category to the input units and observing the activation of the predicate unit. The activation rule is

$$a_X(C/P_1,...,P_j) = \frac{W(P_1,...,P_j) \cdot F(C)}{|F(C)|^2}$$

which reads: The activation of unit X upon presentation of category C given that premises P_1 to P_j have been encoded equals the dot (or inner) product of the weight vector encoding the premises with the C feature vector, all divided by the squared length of the C vector. When C is a conclusion category, this activation value is the model of argument strength.

The weight vector is a non-linearly derived representation of the premises. The projection of the weight vector onto the conclusion category vector is therefore a representation of the projection of the premise categories onto the conclusion category. It corresponds to the features that the conclusion category has in common with the combined premise categories. Geometrically, the model proposes that argument strength is equal to the ratio of the length of this projection to the length of the conclusion category vector. This is the sense in which argument strength is hypothesized to be proportional to the coverage of the conclusion category's features by the premise categories.

Results using Unfamiliar Predicates

Making use of a simple model of the similarity between categories, the feature-based model can be shown to account for 11 of 12 argument strength phenomena (Sloman, 1991). For example, it accounts for the diversity phenomenon above because more diverse premises tend to cover the feature space better than less diverse ones. Another example is a phenomenon that acts as a boundary condition on diversity, feature exclusion. If a premise category shares few features with the conclusion category, it provides little additional coverage and therefore does not contribute to argument strength even if it is dissimilar to other premises.

The one phenomenon not accounted for by the model is called nonmonotonicity. Sometimes, adding a premise can reduce argument strength. For example, introductory psychology students prefer, on average, the argument "Flies have X, therefore Bees have X" to the argument "Flies have X, Aardvarks have X, therefore Bees have X." One interpretation of this phenomenon is that feature consistency is important; perhaps features that appear in one premise but are inconsistent with other premises are given less weight in the feature-matching process. This idea could be implemented in the feature-based model in several ways. A particularly simple way would entail introducing weight decay. If weights are reduced every time that they are updated, then the representations of features appearing in all but the last premise will have lower values if those features do not re-appear in later premises. In the current example, the strength of both arguments depends primarily on the overlap of the features of flies and bees because aardvarks and bees have so few common features. Because flies and aardvarks also have very few features in common, the weights corresponding to the representation of flies would decay in the second argument and therefore be lower than in the first. The reduced values of the flies' representation would lead us to expect the first argument to be stronger. None of the results that we report below would be affected by this generalization of the model because weight decay would have no effect on the model of single-premise arguments.

The model has been tested quantitatively by correlating its predictions to ratings of argument strength provided by subjects. To obtain the predicted strength of an argument from the feature-based model, the model must be given a featural description of each category appearing in the argument. Such featural descriptions were obtained from feature ratings for a set of mammals collected by Tony Wilkie (cf. Osherson et al., 1991). Varying a single parameter (a cutoff which determined a threshold below which feature ratings were set to 0), correlations of 0.96, 0.97, 0.59, 0.83, and 0.77 were obtained on five different data sets, respectively.

Extending the Model to Familiar Predicates

We define a "nonexplainable" predicate as one which is familiar but for which subjects cannot explain the relation between category and predicate. A nonexplainable argument is one containing nonexplainable predicates. We ran an experiment to test our hypothesis that nonexplainable arguments will be treated in the same way as arguments

involving unfamiliar predicates. Subjects rated the strength of arguments with familiar predicates and, afterward, tried to explain the relations among the various categories and predicates. We evaluate our hypothesis in two ways. First, we expect that we should observe the same phenomena with nonexplainable arguments as we do with those using unfamiliar predicates. Our use of single-premise, specific arguments limits us to two such phenomena, similarity and asymmetry. The similarity phenomenon states that arguments tend to be stronger the greater the judged similarity between the premise and conclusion categories. We therefore test for this phenomenon by examining correlations between argument strength and similarity judgments. The asymmetry phenomenon states that the strength of arguments can be changed by reversing the premise and conclusion categories. We evaluate asymmetry by testing the feature-based model's ability to predict differences between the judged strengths of a set of arguments and their reversed counterparts. Finally, we expect the model to make predictions consistent with subjects' strength ratings for nonexplainable arguments. We test this prediction by examining correlations between the feature-based model's predicted argument strengths and subjects' judgments. We compare the correlations we obtain for explainable versus nonexplainable arguments.

Experimental Procedure

We constructed 16 arguments which we expected to be explainable and, using the same categories, another 16 which we expected to be nonexplainable. By exchanging the premise and conclusion of each argument, we obtained a total of 32 arguments of each kind. An argument was deemed explainable if it seemed that subjects would base their judgments on only a small set of features. For example, we believed that the argument

Collies are susceptible to heat stroke.
Siamese cats are susceptible to heat stroke.

would suggest features like "have fur" while

Wolves sometimes attack their mates.
German shepherds sometimes attack their mates.

would suggest features like "can be fierce." Examples of nonexplainable arguments include

Collies hate salted peanuts.
Siamese cats hate salted peanuts.

and

Wolves have dark tongues.
German shepherds have dark tongues.

Each of two groups of 12 University of Michigan students from Introductory Psychology courses were tested on different sets of 8 explainable and 8 nonexplainable arguments. Two other groups of 12 students were tested on corresponding arguments with premise and conclusion statements reversed. Each subject first rated the likelihood of each of the 16 conclusions on an integral scale from 0 to 10. We refer to these estimates as prior likelihoods. Next, they rated the likelihood of the conclusion given the premise on the same scale. The wording of the likelihood question can be inferred from the following example: "Collies hate salted peanuts. How likely do you think it is that siamese cats also hate salted peanuts?" We refer to these estimates as conditional likelihoods. Next, they were asked to briefly explain each premise and conclusion. They were given some example explanations and were encouraged to provide explanations that were sensible though they need not be true. Subjects were also told that if no possible explanation came to mind, they could skip that statement. They also provided a confidence rating of the validity of their explanations but we will not report these data. Finally, they rated the similarity (from 1 to 7) of each premise category to its corresponding conclusion category.

Results

To verify our assessment of explainability, we counted the number of explanations provided for each statement of each argument (out of a possible 24). For each argument, we averaged the number of explanations given for the premise and conclusion. All arguments which had an average of greater than 18 explanations were labelled "explainable" and all others were labelled "nonexplainable". On this basis, 8 of the arguments that we had expected to be nonexplainable were categorized as explainable and 2 explainable arguments were relabelled as nonexplainable. We thereby ended up with 38 explainable arguments and 26 nonexplainable ones.

Similarity. We found evidence for the similarity phenomenon for both explainable and nonexplainable arguments. Because we were interested in the role of similarity in the transfer of belief from premise to conclusion (the conditional likelihood), without the influence of any spurious correlation between similarity and prior likelihood, we looked at the part correlation between i. similarity judgments and ii. conditional likelihoods with priors partialed out. These correlations were significant for both explainable ($r = .40$, $p < .001$) and nonexplainable

arguments ($r = .43$, $p < .001$). We conclude that the similarity phenomenon does indeed hold for nonexplainable arguments and in fact holds for explainable ones as well.

Asymmetry. The feature-based model predicts that reversing premise and conclusion categories will lead to an argument of more or less strength depending on the relative richness or magnitude of the representations of the two categories. The richness of a representation refers to the extent of featural information that is known about a category. Richness would tend to increase with a category's familiarity and complexity. To see why the model predicts these asymmetries, consider its activation rule. The model of the arguments P therefore C and its reversed counterpart C therefore P have identical numerators ($F(P) \cdot F(C)$; cf. Sloman, 1991), but different denominators. The denominators are the magnitudes of the conclusion categories. Therefore, the model predicts that the strength of the argument with the lower magnitude conclusion category will be greater. For example, people often judge "tigers have X, therefore buffaloes have X" to be stronger than its reversal because, according to Osherson et al.'s (1991) feature ratings, the buffaloes vector has a smaller magnitude than the tigers one. Furthermore, the degree of asymmetry should be directly related to the size of the difference between the magnitudes of the two categories.

To test this prediction, we calculated the magnitude of each category using the feature ratings. Based on these magnitudes, we determined whether an argument or its reversal should be stronger. To measure the actual strength of an argument, we used the mean difference between its conditional and prior likelihood judgments. Each strength measure was weighted by the difference between the magnitudes of that argument's categories. This weight reflects the degree of expected asymmetry. A 2 x 2 analysis of variance with one between-argument factor (the explainability of the argument -- explainable or not) and one within-argument factor (predicted asymmetry -- the argument predicted to be stronger or its reversal) revealed a statistically reliable main effect for the predicted asymmetry, $F(1,30) = 4.42$, $p < .05$. No significant main effect for explainability or for the interaction was observed (both $F's < 1$). Apparently, the model was able to successfully predict not only the direction of the asymmetry for nonexplainable arguments, but for explainable ones as well.

Fit of the model. The feature-based model was fit to the data using the equations and feature ratings described above. Because of the feature-rating method used, ratings tended to overestimate the value of nonsalient features (cf. Sloman, 1991). We therefore varied a cutoff which determined a threshold below which feature ratings were set to 0. The cutoff was varied in small discrete increments. The model's predictions were generated using the cutoff that maximized the correlation between the predictions and the data. The data consisted of the mean difference, for each argument, between each subject's conditional and prior likelihood estimates. Because these means represent a combination of judgments by subjects for whom the argument was explainable and those for whom it was nonexplainable, we do not expect these correlations to be extremely high. The relatively small number of times that subjects failed to provide any explanation prevented us from obtaining reliable likelihood estimates for each argument using only those cases. Nevertheless, the maximum correlation (taken over cutoffs) for nonexplainable arguments was 0.66 ($p < .001$). Notice that this correlation is greater than that obtained between argument strength and similarity ratings. The maximum correlation for explainable arguments was much less (0.36; $p < .05$). The difference between the two correlations was marginally significant ($z = 1.57$; $p = .06$). We conclude that these quantitative tests provide some support for the feature-based model as an account of subjects' judgments of the strength of nonexplainable arguments.

Conclusion

A simple model of property induction, alike in many respects to connectionist models of concept-learning, is consistent with a variety of phenomena in a domain of confirmation -- people's willingness to assert properties of natural-kind categories. Our experiment supports our contention that the domain is larger than previously shown. It includes not only arguments with unfamiliar predicates, but those with familiar but nonexplainable predicates as well.

References

Osherson, D., Smith, E. E., Wilkie, O., Lopez, A., and Shafir, E. 1990. Category-based induction. *Psychological Review* 97:185-200.

Osherson, D., Stern, J., Wilkie, O., Stob, M., and Smith, E. E. 1991. Default probability. *Cognitive Science*, 15: 251-269.

Rips, L. 1975. Inductive judgments about natural categories. *Journal of Verbal Learning and Verbal Behavior*, 14:665-681.

Sloman, S. A. 1991. Feature-based induction. Tech Report No. 40, Cognitive Science and Machine Intelligence Laboratory, Univ. of Michigan.