Consistency is the Hobgoblin of Human Minds: People Care but Concept Learning Models do Not

Dorrit Billman
School of Psychology
Georgia Institute of Technology
Atlanta, GA 30332
(404)894-2349
dorrit.billman@psych.gatech.edu

David Dávila
School of Psychology
Georgia Institute of Technology
Atlanta, GA 30332
(404)894-2349
psg94dd@prism.gatech.edu

Abstract

People may be biased to learn categories which not only capture structure in the environment but organize this knowledge in a manner easy to use in reasoning. Concepts organized to contrast consistently on the same attributes as sister categories within a hierarchy may be particularly useful in guiding induction. We assess whether systems of novel categories organized in this manner were also easier to learn. Supervised concept learning was dramatically easier in the consistent over inconsistent contrast condition. We tested whether several models of concept learning would show sensitivity to consistent contrast, as people did, including assessment of a model designed to use information about consistent contrast, TWILIX. None of the models tested (ALCOVE, rational analysis, and TWILIX) showed much sensitivity to the Consistent/Inconsistent contrast. People may flexibly adjust their learning strategy to capitalize on simple regularities when available, in a manner not incorporated in these concept learning models.

Multiple influences conspire to produce our systems of categories and to produce new learning at the edges of existing knowledge. The structure of the environment is a key influence, as we use categories to refer to types of actual entities and to guide us through the world. Second, a person's activities and goals prioritize those aspects of the environment that support important activities over other information whose value is less clear. In addition, the business of mental life and economies of mental activity are important influences. Concepts are used in reasoning, remembering, and imagining. These mental activities and human mental limitations influence category construction as well.

The present work investigates learning biases, or constraints, that make reasoning tasks more straightforward: hierarchy and consistent contrast. A bias to organize information into set inclusion hierarchies, at least local ones, aids many forms of default and deductive reasoning. A bias for consistent contrast aids inductive reasoning. The experiment reported here investigates consistent contrast. *Consistent contrast* is a relation within a set of categories which are daughters of the same superordinate category. For categories with consistent contrast, the attributes relevant to one category are also relevant to the others in the set. Consistent contrast is the principle motivating *variability bias* in the machine learning model TWILIX (Martin, 1992;Martin&Billman, 1991) and is closely related to Goodman's (1983) notion of projectability (also Shipley, 1993; Russell, 1986; Billman, 1992). If individual, known types of animals are homogeneous with respect to diet, one should be able to generalize that the diet observed for one individual of an unfamiliar kind will be generally true for the kind as a whole (Shipley, 1993). The idea of consistent contrast is linked to hierarchy because it is the hierarchy that provides the set of contrasting categories. Type of diet will not be consistent within, or even applicable to, categories such groups of people or kinds of machines.

Evidence for a consistent contrast has come from induction studies that assessed the generalizations which subjects were willing to make from a single instance of a new category. Macario, Shipley, and Billman (1990) found that children's generalizations from a single instance respected consistent contrast. Learning studies could also assess whether a whole set of novel categories are better learned when they contrast consistently. The present study investigates learning. We compared learning sets of three categories in a Consistent Contrast Condition with learning sets of categories in an Inconsistent Contrast Condition. In the Consistent Contrast Condition the same attributes were important across the set. In the Inconsistent Contrast Condition the same individual categories were regrouped such that different attributes mattered for each of the categories in the contrast set. We predicted that the identical categories would be learned more easily when part of a Consistent than Inconsistent Contrast set.

Experimental Method

Subjects. Fifty students from the Georgia Institute of Technology, 26 in the Consistent and 24 in the Inconsistent Condition participated for extra credit. All had normal color vision

Materials. Stimuli were animated events showing alien animals, moving and vocalizing against a background scene. During the learning phase each trial presented one event and the learner was asked to click the appropriate name for the creature from a set of three labels (*yodlar*,*ralfaz*, and *muntog*). Subjects were given feedback and the correct label was displayed. There were 45 learning trials, 15 from each category. Events were composed from six, three-valued attributes: Sound, Movement, Habitat, Color, Head, Body/Legs.

Table 1: Set 1 Stimuli Schema Used for Human Participants

Consistent Contrast

	Configuration 1	Configuration 2	Configuration 3
Category 1	11 xx xx	xx 11 xx	xx xx 11
Category 2	22 xx xx	xx 22 xx	xx xx 22
Category 3	33 xx xx	xx 33 xx	xx xx 33

Inconsistent Contrast

	Configuration 1	Configuration 2	Configuration 3
Category 1	11 xx xx	xx 11 xx	xx xx 11
Category 2	xx 22 xx	xx xx 22	22 xx xx
Category 3	xx xx 33	33 xx xx	xx 33 xx

Stimuli for the three categories for each of six subject groups are shown schematically. Numbers indicate the value of an attribute that was consistently assigned to members of a given category. X's indicate random attribute values for members of the category. Each column indicates an attribute, ordered as sound, movement, habitat, color, head-type, and body/leg.

Each category was defined by a combination of two of the six attributes as shown in Table 1. In the Consistent Contrast Condition, the same two attributes determined category membership for all three categories. For example, in Configuration 1 all *yodlars* croaked and flew, *ralfazes* bleated and walked, while *muntogs* roared and jumped. Three different pairs of attributes were used in three different configurations to counterbalance the effects of attribute salience. In the Inconsistent Contrast Condition, however, each category used a different pair of attributes to mark category membership, for example, Category 1 used sound and movement, Category 2 used habitat and color, while Category 3 used head and body.

The influence of individual attributes and individual attribute values was counterbalanced to equate the impact of these factors in Consistent and Inconsistent Conditions. As shown in Table 1 nine categories were used across the three configurations. These nine individual categories were identical between the Consistent and Inconsistent Conditions, but they were grouped into different category sets. Finally, the identical set of instances are used in a given category (i.e. the xx xx 22 category, 33 xx xx category, etc.) when it occurs in the Consistent or in the Inconsistent Condition; e.g., the identical set of xx 33 xx items are used in Consistent Condition Configuration 2 and in the Inconsistent Condition Configuration 3. A single order of instances was used for all subjects in a given condition and configuration.

The test phase consisted of 30 trials. Half of the test items were normal examples of the three categories seen during learning and half were incorrect events. An incorrect test item scrambled up the assignments of the defining attribute pair (e.g., 12 xx xx rather than the correct 11 xx xx). No labels were provided.

Procedure. Participants worked in sound-isolated cubicles. Subjects were instructed that they would be touring the Saturn zoo and observing different animals. For each, three names would appear and they should click on the name they thought was the correct label for the display. After the subject's judgment, the name the zookeepers use would be displayed. Subjects were told they would be tested later. During the learning phase, subjects saw an animated scene, clicked on one of three category labels, and got feedback indicating the correct category choice.

At the beginning of the test phase, subjects were told they would see more events, some of which would be like what they had seen and some of which would be new. They were told to click a 'yes' button if the event was "like something you had seen before" and 'no' otherwise. Correct, familiar displays were consistent with the schema for the three categories used during learning. Incorrect or discrepant displays disrupted the pairings between diagnostic attribute values which had held during learning. No feedback was provided. Finally, participants filled out a questionnaire about what they noticed.

Design. The independent variables were Condition and Configuration, nested within Condition. The dependent measures were the number of correct classifications over the learning trials and the number correct on the test.

Experimental Results

Average number correct over learning was 86.9% in the Consistent and 49.5% in the Inconsistent Condition. Subjects in the Consistent Condition jumped to high, asymptotic classification in the first 10 trials. The effect of Condition, $F(1,44)=269$, $p<.001$, but not Configuration, $F(4,44)=2.27$, was highly significant. Subjects in the Consistent Contrast Condition also performed dramatically better, mean of 75.1% correct, than those in the Inconsistent Contrast Condition, mean of 55.3%, on test events (Condition $F(1,44)=14.03$, $p=.001$; Configuration $F(4,44)=.87$).

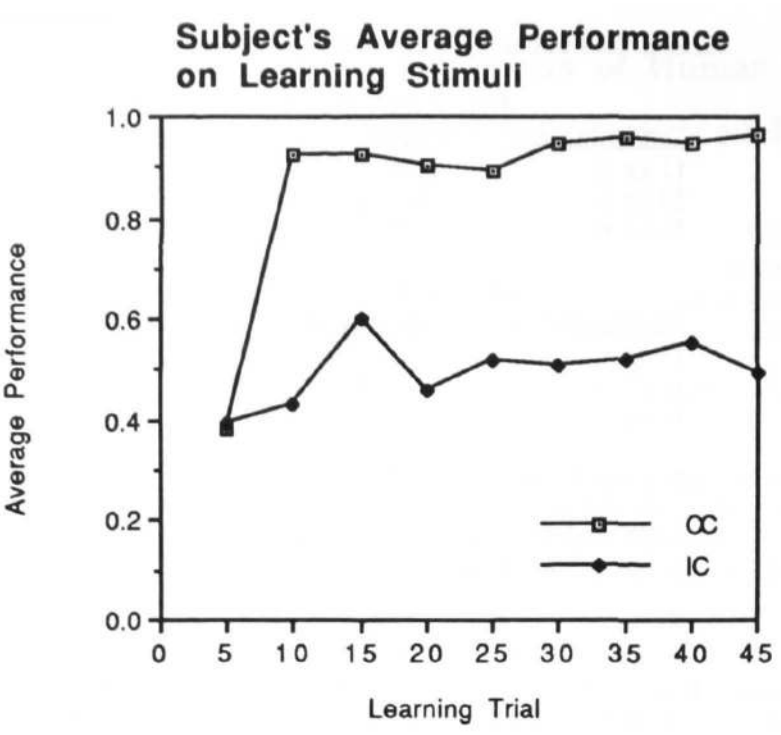

Figure 1: Learning curve for Consistent and Inconsistent Conditions

Methods for Comparative Simulations

We believe these results pose difficulties for many computational accounts of concept learning, including recent modeling that motivated these experiments. We ran simulations on RA (Anderson, 1991), ALCOVE (Kruschke, 1990, 1992), and TWILIX (Martin,1992; Martin&Billman,1991), to compare performance in the consistent to inconsistent condition.

Dependent Variables. All models estimate the probability of each alternative category label and then pick the value with the highest probability to generate their classification response. We use the probability estimates rather than the coarser measure of percent correct responses.

Set 1 Stimuli. We used stimuli from the same schematic specification as shown in Table 1. For the simulations, we also controlled the between category similarity and hence confusibility between categories, as well as within category similarity. In particular, we constructed stimuli (Table 2) such that any advantage of the Consistent Condition could not be due either to greater within-category or lower between-category similarity. Values of the four unpredictive attributes were assigned to make the between-category similarity in the Consistent Condition HIGHER, and hence the categories more confusible, than in the Inconsistent Condition. A model dominated by similarity would learn faster in the Inconsistent than the Consistent Condition. Our purpose was simply to ensure that any advantage for the consistent condition could not be due to simple differences in similarity relations. For all models we averaged ten runs with different stimuli orderings.

Set 2 Stimuli. We used the identical stimuli seen by subjects in the experiment (represented as numbers, not body-parts, of course). The six ordered sets of instances, for the two conditions by three configurations, produced six runs.

Table 2: Set 1 Exact Items Used in Simulations

Consistent Condition Learning			Inconsistent Condition Learning		
Category 1	Category 2	Category 3	Category1	Category 2	Category 3
11 11 11	22 22 11	33 13 12	11 11 11	22 22 11	22 11 33
11 13 12	22 31 21	33 33 22	11 13 12	23 22 12	23 13 33
11 31 21	22 22 23	33 22 23	11 31 21	32 22 21	32 31 33
11 33 22	22 32 33	33 32 33	11 33 22	33 22 22	33 33 33

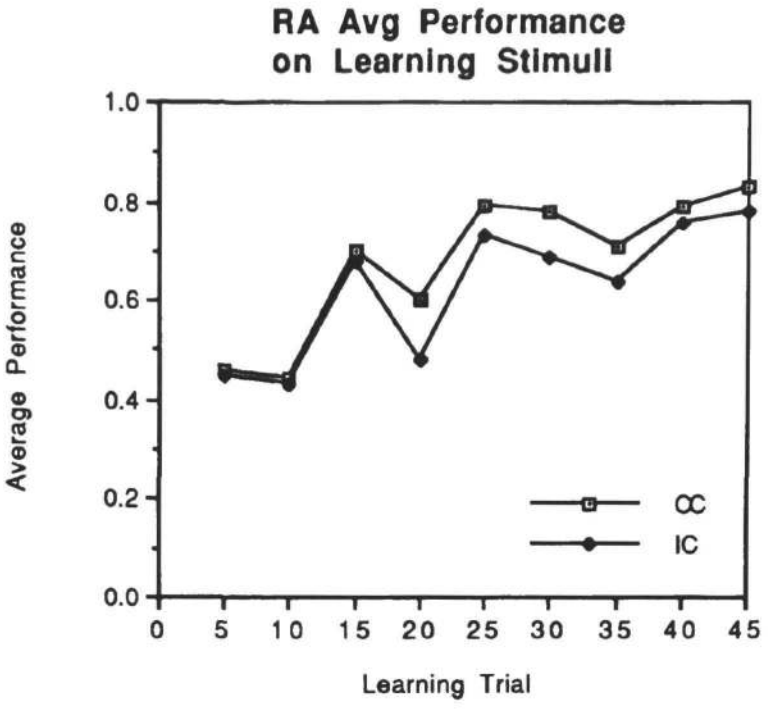

Figure 2: RA's Performance on Set 2 Stimuli

Results from Comparative Simulations

RA. The rational analysis model is a nonhierarchical unsupervised clustering algorithm that approximately optimizes the predictive utility of the clusters it creates. It can be applied to supervised learning by exclusively looking at the model's predictions of the attribute specifying the category label[1]. We anticipated that RA would be insensitive to the difference between the consistent and inconsistent conditions, as RA is indifferent to the basis for predictive success and hence to whether instances in contrasting categories are similar in the same or in differing respects. RA was implementation from Anderson (1991) and run with c set to .3. For the similarity-controlled stimuli of Set 1, RA showed a slight (.04) consistent advantage for the *Inconsistent* Condition over the Consistent throughout the learning curve. After three presentations of the 12 learning items probability estimates of the correct category label were .82 for the Consistent and .86 for the Inconsistent condition. Figure 2 shows RA's probability estimates for Stimulus Set 2, the exact stimuli seen by our subjects. Each of the two curves comes from three runs of RA. Here RA does show a modest but fairly consistent advantage for the Consistent Condition run, though nothing like the strong, early contrast shown by people.

Qualitative-Attribute-ALCOVE. ALCOVE is an instance-based supervised concept learning algorithm with attentional learning. As well as storing instances and generalizing based on the similarity of a novel instance to known instances, it also "stretches" or "shrinks" dimensions of the representation space to learn to weight more heavily those attributes which discriminate between categories. ALCOVE is designed for continuously valued attributes, but can easily be applied to binary categorical attributes by using only 1 and 0. However, application to multi-valued categorical attributes (red/blue/yellow), required code modification. Instead, we linked together sets of attribute values to a single attentional weight representing the attribute as a whole. Each such weight was increased or decreased in accord with the success or failure of each classification. We used code provided by Kruschke modified only as described here.

Since the similarity relations worked to make the Inconsistent Condition categories more discriminable from each other, this should push QA-ALCOVE toward better learning in the Set 1 Inconsistent Condition. However, attentional learning could benefit the Consistent but not the Inconsistent Condition. ALCOVE has a large parameter space. As a result it is strong at fitting a variety of data, but deriving predictions from ALCOVE is difficult. Specifically, it is very difficult to provide any proof of ALCOVE's insufficiency without demonstrating an exhaustive search of its parameter space. This we did not undertake. We did

[1] RA can also be adjusted for supervised learning by changing the system's prior belief that there will be a unique value of the label attribute for each cluster that RA creates. Auxillary runs of RA with this label sensitivity were similar. Runs with the coupling parameter set to .2 produced similar results to those reported here, as well.

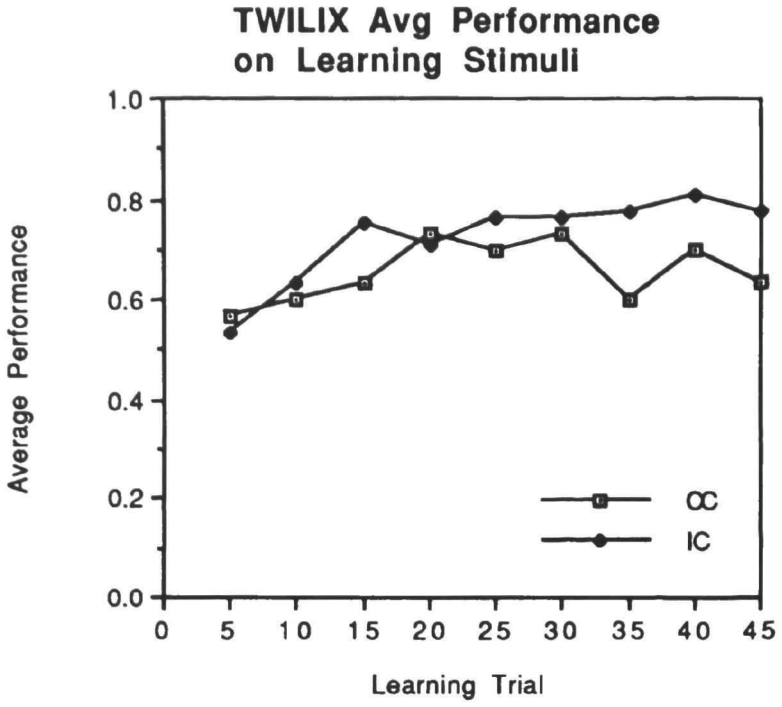

Figure 3: TWILIX's Performance on Set 2 Stimuli

think it informative to see if ALCOVE naturally produced the strong human advantage for the Consistent Condition. We used parameter values from Kruschke (1990, Figs 2.3-2.10) selected to demonstrate the power of ALCOVE's attentional learning, for task and stimuli of broadly similar complexity. All these runs can show is whether for one sensible parameter choice ALCOVE shows any hint of the strong human bias for the consistent categories. We ran ALCOVE on four blocks of Set 1 Stimuli. For performance on the third block, mean probability estimate for the correct label was .360 in the Inconsistent and .361 in the Consistent Condition; number correct on this block averaged 7.8 of 12 for Inconsistent and 7.4 for Consistent. While the conditions were indistinguishable in means, the Inconsistent Condition had slightly greater variability. We did not run ALCOVE on Stimuli Set 2.

TWILIX is a recursive hierarchical clustering algorithm, that includes variability bias. Variability bias alters probability estimates of the consistency of one attribute value within one category using information about the consistency of (other values of) that attribute in contrasting categories. If color is highly consistent within each of several familiar types of jewels, the system is biased to expect that a new type of jewel will also have a characteristic color. TWILIX with its variability bias has been run on an induction task for which human data is available (Nisbett, Kranz, Jepson,&Kunda, 1983), and its pattern of performance (Martin & Billman, 1992) was quite similar to that of people. We anticipated TWILIX would learn faster in the consistent than in the inconsistent condition. Like attentional learning, variability bias is a method of biasing the learner to treat some attributes as more important than others. Unlike attentional learning, variability bias is not a global filter, peripherally screening out information about a given attribute for any purpose in any input. Rather, variability bias is local to a particular context of contrasting categories: while color may be important for types of jewels but not types of cars, it will still be noticed for both types of stimuli. Either way of learning attribute importance could aid learning in the consistent condition, where the same attributes are important across all three categories. In tests on Set 1 Stimuli, TWILIX learns quickly and identically in Consistent and Inconsistent Conditions. At the end of the first block of 12 instances, performance averaged .94 in the Consistent and .97 in the Inconsistent Conditions. From trial to trial the condition advantage switchs, but the average across runs and across trials 1-12 is .77 for the Consistent and .80 for the Inconsistent.

For Stimulus Set 2 in Figure 3, TWILIX also looks decidedly inhuman. Difference between conditions is very small, late, and again favors the *Inconsistent* Condition. TWILIX too is basically indifferent to the contrast between Consistent and Inconsistent Conditions.

Understanding this absence of benefit prompts a closer look at how TWILIX uses variability bias. The largest influence of variability bias will be on the first use of a category. In particular, it will guide the system in when to set up a new category. Criteria for category formation will be more important in unsupervised than supervised learning tasks. In addition, effect of the prior probabilities provided by contrast categories will be quickly tempered as evidence about the category is collected. Thus most of the influence of variability bias will be seen in first setting up a category

(e.g. induction from a single instance) and in estimates of the proportion of category members which have the most frequent attribute value, rather than on ongoing accuracy in predicting the correct value.

Both TWILIX and ALCOVE provide a way of prioritizing some attributes over others, but this sensitivity is too modest to produce the dramatic difference in conditions which people exhibit. In a supervised learning task, when attributes that are reliably informative about all categories are available, people's use of this information apparently swamps sensitivity to other aspects of the problem. This extreme focus or selectivity apparently true for people does not characterize the more "optimal" models presented here. The internal feedback model (Billman & Heit, 1988) is a fourth concept learning model applicable to this task which we did not test on these stimuli. It might learn differently between conditions because it has both strong attentional learning and strong attention limits.

Conclusions

We have found a dramatic difference in the difficulty people have in learning a set of three novel categories which consistently contrast on the same attributes versus a contrast set in which different attributes matter for different categories. Empirically we need to determine to what extent this advantage is due a general shift in periferal attentional versus more strategic knowledge about relevance of certain attributes to certain types of categories. Experiments are in progress that assess consistent contrast in hierarchically organized categories.

Theoretically, our finding poses a challenge to most computational models of concept learning. Most models do not capitalize on noticing and using simple regularities where they exist, and do not predict dramatically easier learning when extensive detail about multiple attributes need not be preserved. To accommodate instance and attribute driven learning may require models that adjust their strategies (Kruschke & Erickson, 1994) or representation in response to the task.

If there is an 'easy' classification rule, people will discover and use it. But what makes a rule easy or hard? One source of ease or difficulty stems from how simple components can be organized into a system of categories useful for deductive and inductive reasoning, as well as capturing accurate, relevant information about the world.

Acknowledgements

Thanks to Angel Cabrera, John Kruschke, Joel Martin, and Terry Shikano for providing code and aid in running the simulations.

References

Anderson, J.R. (1991) The adaptive character of thought. Erlbaum Press: Hillsdate, NJ.

Billman, D. (1992). Modeling category learning and category use: Representation and Processing. In Percepts, Concepts, and Categories: The Representation and Processing of Information. B. Burns (Ed.) Elsevier Science Publishers.

Billman, D., & Heit, E. (1988) Observational learning from internal feedback: A simulation of an adaptive learning method. Cognitive Science, 12, 587-625.

Goodman, N. (1983). Fact, Fiction, and Forecast. 4th edition. Harvard University Press: Cambridge, MA.

Kruschke,J.K. (1990) ALCOVE:An connectionist model of category learning. Cognitive Science Tech Report.#19, Indiana University.

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

Kruschke, J.K. & Erickson,M.A. (1994). Learning of rules that have high-frequency exceptions: New empirical data and a hybrid connectionist model. In Proceedings of the Sixteenth Annual Conference of the Cognitive Science Society. A.Ram and K.Eiselt (Eds.) Erlbaum Press: Hillsdale, NJ.

Macario, J.F., Shipley, E.F., & Billman, D.O. (1990). Induction from a single instance: Formation of a Novel Category. Journal of Experimental Child Psychology, 50,179-199.

Martin, J.D. (1992) Direct and indirect transfer: Investigations in concept formation. Technical Report, Atlanta, GA: Department of Computer Science, Georgia Institute of Technology.

Martin, J. & Billman, D. (1991) Variability Bias and Category Learning. In Proceedings of the Eight International Workshop on Machine Learning L.A Birnbaum & G.C. Collins (Eds.) Pp 90-94, Morgan Kaufman: San Mateo, CA.

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

Russell, S.J. (1986). Preliminary steps toward the automation of induction. In AAAI-86 Fifth National Conference on Artificial Intelligence. Philadelphia Pa: American Association for Artificial Intelligence.

Shipley, E.F. (1993). Categories, hierarchies, and induction. In The Psychology of Learning and Motivation, D.Medin (Ed.). Academic Press: New York, NY.