In D. Hilton (Ed.), Contemporary science and natural explanation: Common sense conceptions of causality. New York: New York University Press.

9 Relationships Between Similarity-based and Explanation-based Categorization*

William D. Wattenmaker, Glenn V. Nakamura and Douglas L. Medin

I. INTRODUCTION

Why do we have the categories we have and not others? This very old question about the structural basis of concepts and categories has recently received renewed attention. In this chapter we approach this issue by asking what makes categories psychologically cohesive.

Traditional answers to the question of what makes categories coherent have relied on the notion of similarity. The consensus has been that concepts should be analysed into constituent components or features, that patterns of matching and mismatching features across concepts provide a metric of similarity, and that similarity relationships constitute the structural basis of categories. For example, *robins* and *eagles* are placed in a different category from *collies*, because *robins* and *eagles* share a number of properties (e.g., wings, two legs) not true of *collies*. Except for the proviso that properties be perceptible to humans, the essential claim of similarity-based approaches to categories is that we have the concepts we have because that's the way the world is structured.

The success of similarity-based approaches to concepts is well documented. Work of Rosch and others arguing that our concepts mirror the correlational structure of the environment has led to a revolution in our view of concepts and categories that has pervaded not only psychology (see Medin and Smith, 1984; Mervis and

* The research described in this paper was supported in part by NSF Grant BNS 84-19756 and by National Library of Medicine Grant LM 04375. We thank Lawrence Barsalou, Dedre Gentner, Denis Hilton, Andrew Ortony, Larry Rendell, and Brian Ross for helpful comments and discussions relevant to this paper.

Rosch, 1981; Smith and Medin, 1981, for reviews) but also linguistics, philosophy, anthropology, and artificial intelligence. The idea that concepts have fixed definitions comprised of singly necessary and jointly sufficient features has been superseded by the view that many concepts are structured in terms of characteristic rather than defining features.

Despite its impressive successes as an account of the structural basis of categories, the similarity-based approach to concepts has recently been severely criticized (Murphy and Medin, 1985; see also Medin and Wattenmaker, 1987; Wattenmaker, Dewey, Murphy and Medin, 1986, for related arguments). We believe that many of these criticisms are well-founded, so much so that we see no way to salvage the view that concepts are structured solely in terms of similarity. The first part of this chapter will briefly describe the similarity-based approach to conceptual structure and then highlight certain problems or shortcomings that we believe are intimately linked to this view.

The upshot of our criticisms of the similarity-based view of concepts is that the focus on the structure of the environment comes at the cost of neglecting the nature of the organism which develops and uses concepts. Part of the answer to why we have the categories we have is that we are the sort of organisms we are. This leads to an alternative approach to concepts which we refer to alternatively as the knowledge- or explanation-based approach to conceptual structure. This approach emphasizes that people's prior knowledge and theories about the world serve to structure concepts and provide *explanations* for *why* certain instances are members of a concept. This explanation-based approach to concepts is in keeping with the spirit of other chapters in this volume that in one way or another address themselves to alternative conceptions of causal explanation.

In the second main section of this chapter we discuss how an explanation-based approach to concepts handles many of the problems associated with similarity-based approaches. We will argue that peoples' theories serve to constrain which properties or features of a concept are selected or made salient, provide the basis for both intra- and inter-conceptual relational structures, and guide inductive inferences and mental simulation.

Assuming that the above arguments can be made in a compelling manner, one way in which this chapter might be read is that there are two approaches to conceptual structure and that the

preponderance of evidence favours one of them (the explanation-based one). But things are not so simple. Many of our criticisms of similarity-based induction, namely that the notion of similarity is too unconstrained, seem equally applicable to explanation-based induction. If there are no constraints on theories, then to say that concepts are organized in terms of theories may be circular. The final sections of this chapter take a second look at similarity and a second look at explanation. We shall argue that there are alternative conceptions of similarity, more structural in character, that provide a basis for integrating similarity- and explanation-based approaches.

The major theme of this chapter is that the road to future progress in understanding human conceptual behaviour lies not with the nature of the world alone nor with the nature of human beings alone, but rather with the *relationship* between intelligent organisms and their environment. This view leads naturally to an interleaving of the notions of similarity and explanation. Similarity and explanation may be mutually constraining, we argue, and there may be some intriguing parallels between similarity-driven and explanation-driven induction. Before developing this more integrated view, however, we need to provide more by way of background.

II. THE SIMILARITY-BASED APPROACH TO CATEGORY STRUCTURE AND CONCEPTUAL COHERENCE

The most pervasive and intuitively plausible explanation of conceptual coherence is that objects, events or entities coalesce to form concepts because they are similar to each other. In this approach, similarity relations among objects in the environment determine the structure of concepts, so that similar objects are placed in the same class and dissimilar objects are placed in different classes. This similarity structure among external objects or events yields distinct clusters, and concept formation is simply a process of internalizing these natural discontinuities.

Although the consensus has been that similarity provides the *metric* for structuring categories, and that similarity can be defined in terms of matching and mismatching properties, there is disagreement concerning which aspects of similarity underlie category structure.

A. The classical view

One view of category structure (inherited from Aristotle) is that natural language concepts are characterized by simple sets of defining features that are singly necessary and jointly sufficient to determine category membership (e.g., Katz and Postal, 1964). A candidate exemplar either does or does not possess these defining features and thereby is or is not a member of the category. The major problem with the classical view is that research suggests that the majority of natural concepts are not organized around defining features but rather are structured in terms of sets of typical or characteristic features (see Medin and Smith, 1984; Mervis and Rosch, 1981; Smith and Medin, 1981, for recent reviews).

B. The probabilistic view

The rejection of the classical view of categories has been associated with the ascendance of the probabilistic view of category structure (Wittgenstein, 1953). The current consensus has it that categories are "fuzzy" or ill-defined, and that they are organized around a set of properties or clusters of correlated attributes that are only characteristic of category membership. Membership in a category can thus be graded, rather than all-or-none, where the better members have more characteristic properties than the poorer ones. In an attempt to be specific about the structural basis of graded categories, Rosch and Mervis (1975) had subjects list properties of exemplars for a variety of concepts such as *bird*, *fruit* and *tool*. They found that the listed properties for some exemplars occurred frequently in the concept while others had properties that occurred less frequently and, most importantly, the more frequent an exemplar's properties were, the higher its ratings for typicality in that category. Rosch and Mervis formalized the notion that categories are organized by a principle of *family resemblance*. They developed a measure for the prototypicality of an example that increases with the frequency of the properties it shares with members of its own category and decreases with the frequency of properties it shares with members of contrasting categories (cf. Tversky, 1977, pp. 347–9). Less formally, family resemblance increases with within-category similarity and decreases with between-category similarity. Family resemblance is highly correlated with the speed with which an exemplar can be

categorized as well as with other typicality effects (see Rosch and Mervis, 1975).

In related work, Rosch and her associates (Rosch, Mervis, Gray, Johnson and Boyes-Braem, 1976) found that one level of abstraction, which they call the *basic level,* is more fundamental than either more specific subordinate categories or more abstract superordinate categories. For example, by their criteria, *chair* would be a basic level concept, but *furniture* and *rocking chair* would not be. These claims are reinforced by a variety of empirical results (see Mervis and Rosch, 1981). For example, basic level categories are the ones first learned by children, most likely to be shared by people of different cultures, and most rapidly classified in reaction time experiments.

There have been numerous attempts to be more specific about the structural underpinnings of basic level categories. For example, Rosch *et al.* suggested that basic level categories maximize both component cue validity (the probability that an entity belongs to category j given that feature i is present) and within-category similarity relative to between-category similarity. Neither claim stands up to closer scrutiny. Cue validity is maximized for the most abstract or general categories and it is impossible to simultaneously maximize within-category similarity and minimize between-category similarity (Medin, 1983; Murphy, 1982). These problems notwithstanding, the ability of the family-resemblance (or probabilistic) view to address findings that are problematic for the classical view, coupled with converging operations that reinforce the notion of a basic level, makes a good case for the idea that our categories mirror the correlational structure of the environment.

Despite considerable support for the probabilistic view, we believe that evidence from two major sources reveals that the probabilistic view is fatally flawed. The first source of evidence is primarily empirical and grows out of our attempts to be more specific about the structural basis of family resemblance categories. Here the central problem is the practice of equating concepts with lists of independent attributes or features. The second source of evidence is mainly theoretical or conceptual and it questions the fundamental notion of similarity. Our criticisms apply not only to the Rosch and Mervis operational definition of family resemblance but, more generally, to all current similarity-based approaches to conceptual coherence (see Murphy and Medin, 1985).

1. Empirical problems for the similarity approaches to family resemblance categories

a. *Family resemblance sorting.* According to a family resemblance principle, categories are organized around exemplars that are prototypical of potential categories. In Rosch's words, the idea is "that potential prototypes will tend to become centers of categories in free sorting" (Rosch, 1975, p. 196). That is, if we construct artificial categories by selecting prototypes and generating examples to create a family resemblance structure, then these same categories should be reproduced when people are allowed to construct their own categories from these examples. This prediction is a natural consequence of viewing concepts as comprised of sets of independent features.

The idea that people will prefer to sort entities into categories organized around a prototype was examined in a recent set of studies in our laboratory (Medin, Wattenmaker and Hampson, 1987). Figure 1 presents an abstract description of a set of 10 entities and two alternative means by which they might be sorted into two equal-sized categories. The dimensions correspond to types of components or features and 1 and 0 correspond to values on these dimensions. For example, D1 might be colour and 1 might correspond to a red stimulus and 0 to a green stimulus. The abstract notation, 1111, might correspond to a stimulus consisting of one large red triangle and the notation, 0000, to a stimulus consisting of two small green circles.

The sort on the bottom left side of Figure 1 is labelled as family resemblance and the topmost example in each category represents the prototype or best example of the category. Each of the other examples would match the best example on three of its four values. An alternative sorting strategy is to partition the examples on the basis of values on a single dimension (in the example in Figure 1, the first dimension, or colour). If all components are roughly equal in importance, it is easy to see that the family resemblance sort maximizes the average within-category similarity minus the average between-category similarity. For the family resemblance partitioning there is an average of 9.6 within-category matches (the first example has 12 matches to other category members and the other four examples have 9 matches to other members; self matches are not counted) and an average of 6.4 between-category matches, yielding a difference of 3.2 matches (mismatches will show a mirror-

DIMENSION

Example	D_1	D_2	D_3	D_4
1	1	1	1	1
2	1	1	1	0
3	1	0	0	0
4	0	1	0	0
5	0	1	1	1
6	0	0	0	0
7	0	0	1	0
8	1	0	1	1
9	0	0	0	1
10	1	1	0	1

Family Resemblance Sort								*One-Dimensional Sort*							
Category A				Category B				Category A				Category B			
Dimension				*Dimension*				*Dimension*				*Dimension*			
D_1	D_2	D_3	D_4	D_1	D_2	D_3	D_4	D_1	D_2	D_3	D_4	D_1	D_2	D_3	D_4
1	1	1	1	0	0	0	0	1	0	0	0	0	1	0	0
1	1	1	0	0	0	0	1	1	1	1	1	0	1	1	1
1	1	0	1	0	0	1	0	1	1	1	0	0	0	0	0
1	0	1	1	0	1	0	0	1	1	0	1	0	0	1	0
0	1	1	1	1	0	0	0	1	0	1	1	0	0	0	1

Figure 1

Abstract notation for the ten examples from the sorting experiments. D_1, D_2, D_3 and D_4 refer to component dimensions and the values 0 and 1 represent the two alternative features (e.g., red versus green in the dimension of colour). The partitioning on the bottom left represents a sort consistent with family resemblance principles, and the partitioning on the bottom right represents a sort consistent with the use of a single dimension.

Note: From "Family resemblance, conceptual cohesiveness, and category construction" by D. L. Medin, W. D. Wattenmaker and S. E. Hampson, 1987, *Cognitive Psychology, 19*, p. 246. Copyright 1987 by Academic Press, Inc. Reprinted by permission.

image pattern and can be ignored in this example). For the one-dimensional partitioning there is an average of 9.2 within-category matches and an average of 7.2 between-category matches, yielding a difference of 2.0 matches. Comparing the two sorting strategies it is clear that the family resemblance partitioning produces greater within- relative to between-category similarity for the situation where the constituent dimensions are equally weighted.

The general procedure involved asking participants to examine the stimuli carefully and place them into two equal-sized groups in "a way that seems natural or sensible." The exact stimulus material employed varied from study to study. Figure 2 illustrates one such set. The cartoon-like animals map directly on to the abstract structure in Figure 1. For this particular set the exact realization of some value (e.g., "striped") varied from animal to animal. The figure shows the animals grouped by what would be a family resemblance sorting with the respective prototypes for each group clustered in the centre.

The results of our sorting studies are easy to describe. We failed to find any evidence whatsoever that people construct categories that have a family resemblance structure. Across a variety of different stimulus materials, instructions, category structures and task demands we never observed family resemblance sorting. Instead, people showed a strong preference for unidimensional sorting despite our varied efforts to prevent subjects from sorting on the basis of a single dimension. Even when we took measures to prevent unidimensional sorting by constructing stimuli that had three values on each dimension and requiring people to sort the examples into two categories, we still did not observe family resemblance sorting.

A possible response to these results is to argue that, although family resemblance categories do not emerge in sorting tasks, in learning situations family resemblance structures are more natural. As we shall see, however, learning studies reveal parallel findings.

b. *Linear separability.* One way to conceptualize the process of classifying stimuli on the basis of similarity to prototypes is that it involves a summing of evidence against a criterion. If an instance has a criterial sum of weighted properties it will be classified as a bird, and the more typical a member is of the category the quicker the criterion will be exceeded. The key aspect of this prediction is

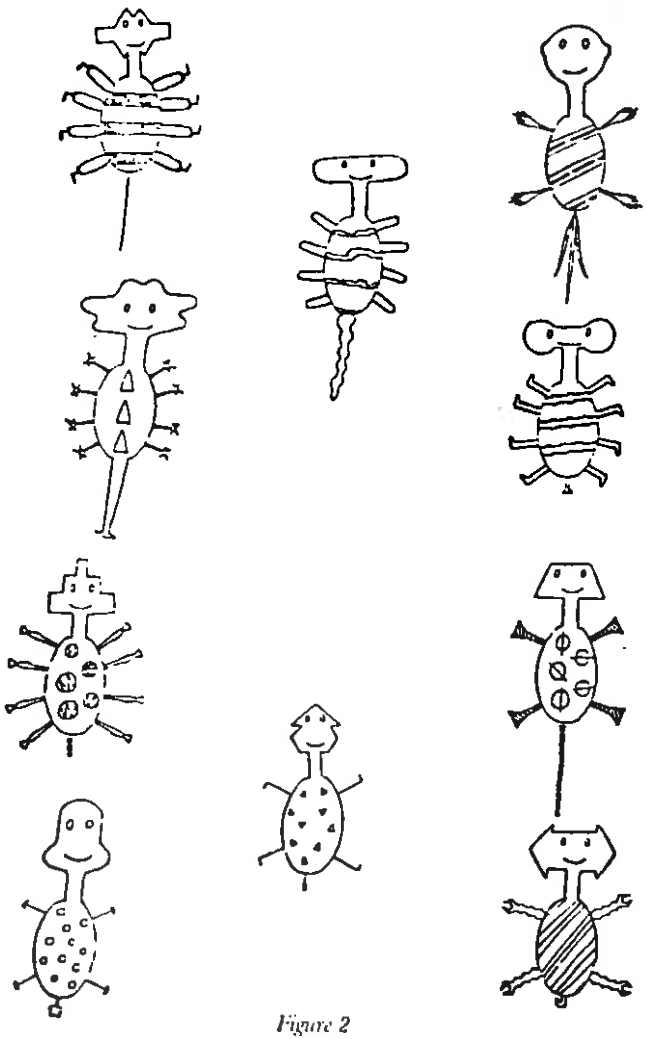

Figure 2

One instantiation of the abstract structure indicated in Figure 1. The four dimensions are body markings (spot vs. stripes), head shape (round vs. angular), tail length (short vs. long), and number of legs (4 vs. 8). The drawings are grouped by a family resemblance principle.

Note: From "Family resemblance, conceptual cohesiveness, and category construction" by D. L. Medin, W. D. Wattenmaker, and S. E. Hampson, 1987, *Cognitive Psychology, 19*, p. 250. Copyright 1987 by Academic Press, Inc. Reprinted by permission.

that there must exist some additive combination of properties and their weights that can be used to correctly partition instances into members and nonmembers.

There is a formal similarity between this constraint and linear discriminant algorithms used in machine pattern recognition (e.g., Nilsson, 1965). If exemplars from two different categories have n properties, then the categories are said to be linearly separable if one can find a set of weights for the n properties such that a linear discriminant function yields higher values for all instances of one category than for those of the other category (Sebestyn, 1962). For a prototype process to work in the sense of accepting all members and rejecting all nonmembers, the categories must be linearly separable. Technically, categories that are not linearly separable should be impossible to learn according to this perspective. However, given that subjects may employ auxiliary processes (e.g., memorization) when confronted with repeated failure, these theories can be interpreted as only making the weaker claim that non-linearly separable categories should be very difficult to learn.

Although linear separability is an important constraint in formal models of classification, there has been very little work attempting to see if linear separability is a meaningful constraint on human classification learning. Of course with natural categories this constraint cannot be examined unless one can ascertain the underlying components of properties of exemplars. To avoid the ambiguities associated with specifying the components of natural categories, researchers interested in structural constraints resort to constructing artificial categories where presumably the component properties can be specified to create the category structures that are of interest.

Figure 3 illustrates categories that are or are not linearly separable. The stimuli consist of values on four components described in terms of a binary notation. The essential difference between the *LS* (top) and the *NLS* (bottom) categories is that the *LS* categories can be separated on the basis of characteristic features. Every Category *A* member and no Category *B* member has three of the four characteristic values for Category *A* (value 1). Thus each exemplar in both categories could be correctly classified by summing the typical values. If the exemplar contains three out of four typical values for Category *A* then it is a member of Category *A*; if the exemplar contains less than three typical values of Category *A*

Linearly Separable Categories

	CATEGORY A DIMENSION					CATEGORY B DIMENSION			
EXEMPLAR	D_1	D_2	D_3	D_4	EXEMPLAR	D_1	D_2	D_3	D_4
A_1	1	1	1	0	B_1	1	1	0	0
A_2	1	0	1	1	B_2	0	0	0	1
A_3	1	1	0	1	B_3	0	1	1	0
A_4	0	1	1	1	B_4	1	0	1	0

Categories Not Linearly Separable

	CATEGORY A DIMENSION					CATEGORY B DIMENSION			
EXEMPLAR	D_1	D_2	D_3	D_4	EXEMPLAR	D_1	D_2	D_3	D_4
A_1	1	0	0	0	B_1	0	0	0	1
A_2	1	0	1	0	B_2	0	1	0	0
A_3	1	1	1	1	B_3	1	0	1	1
A_4	0	1	1	1	B_4	0	0	0	0

Figure 3

Abstract representation of the alternative categorization tasks used in the studies of linear separability. Each task involved eight stimuli varying along four dimensions.

Note: From "Linear separability and concept learning: Context, relational properties, and concept naturalness" by W. D. Wattenmaker, G. I. Dewey, T. D. Murphy, and D. L. Medin, 1986, *Cognitive Psychology, 18*, p. 177. Copyright 1986 by Academic Press, Inc. Reprinted by permission.

then it belongs in Category *B*. A similar algorithm using the number of typical Category *B* values would also lead to unambiguous classifications.

The categories that are not linearly separable have a similar overall distribution of values (Category *A* has five more typical values than Category *B* in both the *LS* and *NLS* cases) but they cannot be partitioned by summing the typical values. For example, exemplar *A*1 has more values characteristic of Category *B* than of Category *A*. Similarly, exemplar *B*3 has more values typical of Category *A* than of Category *B*. A weighted additive function that would correctly partition the category members does not exist. The categories are not linearly separable.

If linear separability is a meaningful constraint on human classification, then people should find the classification task outlined at the top of Figure 3 easier to master than the task at the bottom of Figure 3. To once again make a long story short, however, across a variety of instructions, stimuli and category sizes we were able to find no evidence that linear separability acted as a constraint on people's classification learning (e.g., Medin and Schwanenflugel, 1981).

c. *Related findings.* Probabilistic view theories treat concepts as relatively static and context-independent. The concept *bird*, for example, is thought to be instantiated with the prototype or best example of a bird independent of the prevailing context. Roth and Shoben (1983), however, have shown that typicality judgments vary as a function of particular contexts. For example, college students in the United States judge tea to be a more typical beverage than milk in the context of librarians taking a break, but this ordering reverses in the context of truck-drivers taking a break. In addition, Barsalou (1983) observed that people frequently appear to construct new categories as they become necessary to achieve a current but novel goal. For example, if someone is going on a camping trip for the first time he or she may construct "ad hoc categories" for *places to go camping, things to take camping*, and so on. The ability to construct new categories further illustrates people's highly dynamic and flexible ability to construct representations that meet contextual constraints. Barsalou (1985) also demonstrated the standard typicality effects for ad hoc or goal-derived categories but in this case the underlying basis for typicality was not similarity to a prototype but rather proximity to an ideal value. For example, for the

category, *things to eat on a diet*, typicality judgments were a function of how closely an example conformed to the ideal of zero calories. Again, these observations do not fit with the context-free, similarity-driven abstraction models that characterize the probabilistic view.

2. Conceptual problems for similarity

a. *The flexibility of similarity.* There are a number of reasons to believe that similarity is simply too unconstrained a notion to offer a complete account of conceptual cohesiveness. Consider for example Tversky's (1977) theory of similarity that defines similarity as a function of common and distinctive features weighted for salience. One difficulty with constraining similarity is that in this formulation similarity relations among a set of entities will depend on the particular weights given to individual properties. This is problematic in that Tversky (1977) has demonstrated that the relative weighting of a feature varies with stimulus context and experimental task (see also Gati and Tversky, 1984). Thus there is no unique answer to the question of how similar one entity is to another. To make matters worse, Ortony, Vondruska, Foss, and Jones (1985) have argued that Tversky's model is *too constrained* (!) in that it assumes that a given feature has the same salience regardless of the entity in which it occurs.

A second problem with defining similarity in terms of common and distinctive attributes is that no constraints have been provided for determining what features will be selected. Thus, as Murphy and Medin (1985) pointed out, the number of attributes that *plums* and *lawnmowers* have in common could be infinite: both weigh less than 1,000 kg, both are found on earth, both are found in our solar system, both cannot hear well, both have an odour, both can be dropped, both take up space, and so on. Any two entities can be arbitrarily similar or dissimilar depending on what is to count as a relevant attribute. In fact, Watanabe (1969) offered a formal proof known as "the theorem of the Ugly Duckling" showing that there is no objective justification for preferring any one partitioning of entities over other possibilities. An implication of this proof is that, logically, any two classes are as similar to each other as any other pair of classes.

b. *Attributes as building blocks.* The Rosch and Mervis (1975) measure of family resemblance in terms of matching and mismatching properties is a convenient simplification, and served their purposes rather nicely. This approach implicitly assumes, however, that the properties listed by subjects can be treated as primitives which can be added together independently. However, most concepts are not a simple sum of features (e.g., Armstrong, Gleitman and Gleitman, 1983; Wattenmaker, Dewey, Murphy and Medin, 1986). All the features that are characteristic of a bird do not make a bird—unless these properties are held together in a bird structure. The properties typically listed for the concept bird—laying eggs, flying, having wings, having hollow bones, building nests in trees, having feathers, and singing—each represents a complex concept with both internal structure and an external structure based on inter-property relationships. Building nests is linked to laying eggs, and building them in trees poses logistic problems whose solution involves other properties such as wings, hollow bones and flying. Therefore, it seems that the properties associated with many lexical concepts are anything but an independent set of non-decomposable primitives. A person who simply memorized the attributes of a concept "bird", for example, would have a very odd idea of a bird. Simple listings of independent features lead to a very different *notion* of a bird because information about specific relationships of category features to each other is missing.

A consequence of assuming that categories consist of collections of independent properties or unspecified property relationships is that it does not include information about operations, transformations and relations among properties. In this conception, categorization involves a passive matching of independent attributes to a summary representation, or a passive matching of attribute combinations to an exemplar representation. Rips and Handte (1984) discuss an example where people agree that an object five inches in diameter is more similar in size to a *coin* than to a *pizza* but nonetheless is more likely to be a pizza than a coin. One possible interpretation of these results is that people's real-world knowledge about the size of coins as mandated by law is intruding on the categorization decision. Thus the categorization decision is based on the application of real-world theories to the task, rather than a simple, passive process of attribute matching.

In many instances categorization might involve making a number of inferences and causal attributions. For example, jumping into a swimming pool with one's clothes on is in all probability not directly

associated with the concept *intoxicated*. However, observing this behaviour might lead to classifying the person as drunk. Clearly one's concepts are intimately related to one's theories, and a concept may be instantiated when it can explain an event and is consistent with causal schemata rather than when it matches an object's attributes.

3. Summary

Although there is clear evidence that many natural categories are fuzzy, the similarity-based approach has been unsuccessful in clarifying the structural underpinnings of these findings. Although family resemblance measures derived from attribute listings do predict typicality judgments, there appears to be something fundamentally wrong with family resemblance as a measure of conceptual coherence. When we incorporate family resemblance structures into artifically constructed categories, people do not sort by family resemblance nor does the presence or absence of a structural constraint associated with the family resemblance view, linear separability, have any discernible effect on people's category learning. In addition, there are at least three major problems with similarity-based approaches which treat attributes as independent: (1.) there are no clear constraints on feature weighting associated with concept changes; (2.) similarity will depend crucially on which out of a potentially infinite set of attributes or features are selected; and (3.) relational properties associated with structured descriptions are ignored. Although we maintain that similarity cannot explain conceptual coherence, as will become evident later, we do not wish to imply that a similarity-based approach is completely wrong or useless. Our primary argument is that in order to explain the richness of conceptual structure it is necessary to consider concepts in terms of the broader goals and general world knowledge that the categorizer possesses. We now turn our attention to a more direct discussion of the importance of theoretical and causal knowledge in categorization. Later on, we will have a second, more constructive look at similarity.

III. A KNOWLEDGE-BASED APPROACH TO CONCEPTUAL COHERENCE

Our thesis is that representations of concepts should be thought of as embedded in theories people have about the world. In this section we will attempt to show that viewing concepts in this way can account for many problems that a similarity-based approach fails to achieve.

A. Family-resemblance constructions and theories

1. *Family resemblance sorting.* In an earlier section of this chapter we described a series of failed attempts to obtain family resemblance category constructions. Family resemblance is defined in terms of matching and mismatching independent features, and consequently in these experiments the features that we employed had few clear inter-property linkages. A knowledge-based approach, however, emphasizes the importance of causal and theoretical structures in conceptual coherence and is more consistent with the idea that a network of inter-property relationships links properties to each other and provides conceptual cohesiveness (see Murphy and Medin, 1985, for an amplification of this argument). If this is so, then it may be a mistake to look for family resemblance sortings in contexts where the component properties bear little or no conceptual relationships to each other. To examine this possibility in a second set of sorting studies, we used properties that had salient inter-property relationships (Medin, Wattenmaker and Hampson, 1987). In these experiments we used either trait descriptions that were related to a basic personality dimension (e.g., introversion vs. extroversion) or an occupational stereotype, or alternatively we used cartoon-like drawings of animals where the particular properties could be related to each other in terms of environmental adaptations. In each of these cases some underlying theme or theoretical knowledge exists that can potentially serve to integrate the individual properties. When these materials were used, for the first time we were able to observe category constructions that had an underlying family resemblance structure. These results contrast sharply with the previous experiments, in which family-resemblance constructions were never observed. These findings are consistent with the following claim: categories that have a family resemblance structure may derive their coherence not from overlapping characteristic properties but from the complex web of causal and theoretical relationships in which these properties participate.

2. *Linear separability.* According to the view we have been

developing, whether or not some structural property is important should depend on the type of knowledge structure in which it is embedded and the inter-property relationships appropriate to that context.

What type of inter-property encoding is compatible with linearly separable categories and a summing strategy? Presumably it is important that all the features must be perceived to reflect some common element. For example, consider an object with the following properties: "made out of metal", "has a regular surface", "medium size" and "easy to grasp". Out of context, each of these properties is distinctive, and although one can think of many possible inter-property relationships, no one appears to be particularly salient. Consider, however, the situations where one might be looking for something to use as a substitute for a hammer. In this context each of the properties can be readily linked to the superordinate concept "hammer substitute", and the notion of integrating or summing components to determine overall suitability may become more sensible. On this interpretation, *LS* categories might become easier to learn either because participants would find it natural to add up the number of hammer-like properties or because they might mentally construct a (hammer-like) object from each description and make categorization decisions based on how well the object might serve as a hammer substitute. In fact, Wattenmaker, Dewey, Murphy and Medin (1986) found that presentation of the hammer theme did facilitate learning *LS* categories.

In general, we have found (Nakamura, 1985; Wattenmaker, Dewey, Murphy and Medin, 1986) that learning *LS* categories is facilitated when knowledge structures are activated (e.g., the "hammer substitute" idea) that encourage subjects to encode properties in relation to a superordinate theme and to sum across the dimensions in order to make a categorization decision. Of course, not all interproperty relations or property-theme relations are consistent with a summing strategy. Wattenmaker *et al.* ran other studies where conjunctions of features were made salient by relevant knowledge structures, and under these conditions the learning of nonlinearly separate categories was facilitated. Thus it is not the case that ease of learning can be specified in terms of the configuration of independent features inherent in the category structure. Rather the types of inter-property linkages that are promoted by relevant knowledge, and the compatability between these linkages and the structure of the categories, are the key determinants of learning difficulty.

B. Correlated attributes and causal knowledge

It has often been suggested that conceptual coherence derives from the presence of correlated features. Rosch, Mervis, Gray, Johnson and Boyes-Braem (1976) and Rosch (1978), for example, proposed that natural categories divide the world according to clusters of correlated attributes that "cut the world at its joints." Basic level categories are said to maximize the correlation structure of the environment by preserving these feature correlations. A problem with this correlated-attribute perspective is that there are so many possible correlations that it is not clear how particular correlations are detected (see Keil, 1981). Furthermore, the similarity-based approach fails to distinguish between simple correlations and correlations that derive from underlying causal mechanisms.

To pursue this issue, Medin, Wattenmaker and Hampson (1987) examined the possibility that people would be more likely to construct categories around correlated attributes that could be causally related than around correlated attributes that could not be causally related. For example, in one experiment the stimulus materials were medical symptoms, and subjects could construct categories with symptom pairs that either could be causally-linked (e.g., dizziness and earache) or symptom pairs that would be more difficult to link causally (e.g., sort throat and skin rash). Similarly, in another experiment the stimulus materials consisted of descriptions of animals, and again subjects could base their partitionings on correlated properties that could be causally related (e.g., brightly coloured and poisonous) or correlated properties that were more difficult to relate causally (e.g., long-tailed and slow).

In both experiments people demonstrated a strong preference to sort on the basis of correlated attributes for which a causal or explanatory link could readily be made, and verbal reports indicated that subjects used these causal linkages to justify their constructions. In addition, people chose to sort on the basis of correlated attributes rather than a single dimension only in the case where the correlation was causal in nature. Clearly, causal knowledge plays a fundamental role in conceptual coherence by selecting correlated attributes and in specifying their relationship.

There is also evidence that theoretical expectations can dominate empirical relationships in the perception of correlations. Chapman and Chapman (1967, 1969) presented evidence that therapists and undergraduate subjects perceived correlations between test results and psychological disorders when in fact there were no correlations or the correlations were in the opposite direction. For example, in the draw-a-person test, observers have the belief (part of their personal theories) that suspiciousness of others will be revealed by how the eyes are drawn, and a belief of this sort prevents observers from detecting the empirical correlations. Many of the features we believe to be correlated are probably generated by theoretical knowledge rather than based on observation.

C. Theories and attribute selection in natural concepts

In an earlier section we discussed the difficulties associated with specifying what attributes are selected. We would like to argue that attributes are selected based on people's general world knowledge. For example, one way to investigate the types of attributes that are associated with some concept is to examine the types of attributes that subjects list for a concept. Indeed, this procedure has been employed by Rosch and Mervis (1975) who found that listed attributes can be used to predict accurately both goodness of example ratings and reaction times to verify category membership. Conceivably, the types of attributes people list in this type of a task should be heavily constrained by the relation that a concept has to broader knowledge. That is, rather than listing every possible attribute of a concept, people will provide only those features that are particularly salient as dictated by background knowledge (see Tversky and Hemenway, 1984, for a related discussion). For example, people are unlikely to list flammable as a property of money, not because it does not burn but because flammability is not central to the role money plays in our theories of economic interaction. However, flammability would be more likely to be listed as a property of wood because of the importance of burning wood in human activities.

The central point is that the majority of attributes a concept has are seen as irrelevant and are not even considered (e.g., flammable for money). It is the place of the concept in the network of everyday activities and knowledge that determines whether or not we associate a particular feature with a concept. It seems that

attributes are not in any sense "given" but rather derive from the role of the concept in broader knowledge, functions and goals. In addition, attribute listings tend to leave out important information such as the relational properties that we have argued are important in structuring concepts.

If attribute listings fail to capture the complete structure of concepts, why is it that attribute listings predict goodness of example ratings? One possibility is that both attribute lists and goodness of example ratings are constrained by the same theories. Indeed, Barsalou (1985) found not only that exemplars of goal-derived categories (e.g., things to take out of one's house in case of fire) have typicality ratings that correlate with the degree to which they satisfy the relevant goal, but additionally, natural categories might be organized in terms of dimensions that reflect the interaction between people's goals and activities and the concept. For example, for the concept fruit, "how much people like it" was significantly correlated with exemplar goodness even when the effects of frequency and family resemblance were partialled out.

Although the issue of attribute selection is not important in laboratory studies using impoverished stimulus materials such as red triangles and blue circles, it becomes a dominant factor when one introduces a richer set of stimulus materials. This is amply illustrated by some recent research on rule induction carried out by one of us (Glenn V. Nakamura). The basic task consisted of presenting pre-classified examples taken from more than one category and asking people to come up with a rule that can be used to determine category membership. The category examples were children's drawings associated with the draw-a-person test used in clinical assessment. For a fixed set of drawings, the category labels were varied. For example, one group of participants was told that the drawings were done by mentally healthy versus disturbed children, another group was told that the drawings were done by creative versus non-creative children, and still another group was told that the drawings were done by farm versus city children. The results cannot be described simply in terms of the relative salience of a fixed set of properties, because category labels and descriptive units were not independent. For example, participants in one condition might note that the humans drawn by farm children all had at least some animal parts in them (e.g., a pig-like nose), but

when the same drawings were labelled as creative or mentally disturbed no participant mentioned the presence of animal parts.

These results support an important point based on Kant's distinction between the productive use of imagination (embodying innate perceptual constraints) and the reproductive use of imagination (involving relating particular experiences). Kant's terms (Kant, 1788/1933) correspond closely to Wittgenstein's distinction between "seeing the object" and "seeing the object as." The same object (e.g., a triangle) can be *seen as* a geometric drawing, a wedge, a mountain, a triangular hole, a threat, an arrow, and so on (see also Barresi, 1981). We think this distinction is also important for the draw-a-person stimuli. Actually, in each of the different labelling conditions people stated rules at an abstract level and used the inferred properties of the drawings as support. For example, in the case of drawings supposedly done by farm children the rule might be that "each drawing reflects some aspect of farm life." One drawing might be seen to have a pig-like nose, another to have a farmer's work clothes, and so on. These observations suggest that the drawings do not manifest some fixed set of properties so much as "support" a set of descriptions that derive from the interaction of the drawings with particular observers.

D. Fixed semantic structure and dynamic theories

We have been emphasizing that existing models of categorization have focussed exclusively on similarity relations while ignoring inter-property and inter-concept relations. It should be noted that much of the research in semantic memory based on network models has attempted to represent these relationships. However, this approach has some characteristic shortcomings.

Johnson-Laird, Herrmann and Chaffin (1984) argue that it is difficult to capture the flexibility of human memory in terms of fixed semantic structure. Much of our knowledge is computed by extension to real-world examples rather than pre-stored. To borrow one of their examples, people know that a tomato is more "squishable" than a potato, although it would seem implausible that this fact is directly represented in semantic memory. Additionally, relational properties like "squishable" are highly dependent on context, and the operations of freezing the tomato and boiling the potato would reverse the observations. In order to account for the flexibility that is possible with human memory, it is necessary that

these mental simulations be constrained by world knowledge and people's causal theories about the world. This knowledge allows us to distinguish between transformations that are irrelevant to some relation, from those that are relevant (e.g., boiling versus spray-painting potatoes for the operation of squishing). Theories are dynamic and represent mental models of the world based on perception, memory and imagination.

E. Summary

In this section we have attempted to show that people's theories and knowledge of the world play a major role in conceptual coherence and that many issues that are problematic for a similarity-based approach can be explained by a knowledge-based approach. For example, we argued that categories derive their coherence not from sets of matching or mismatching properties, but rather from a complex interleaving of causal and theoretical relationships in which these properties participate. Causal knowledge works to make certain patterns of correlated properties more salient than others and to provide an underlying basis for their relationship. Furthermore, the role of a concept in everyday knowledge and activities constrains which of a potentially infinite set of properties is associated with the concept. Viewing concepts as embedded in theories appears to answer many of the difficulties associated with similarity-based approaches to concepts.

IV. INTERLEAVING OF SIMILARITY AND EXPLANATION

Despite the ability of the explanation-based approach to handle these problematic issues, many readers perhaps still cling to the intuition that similarity does influence categorization and that much of the time entities are placed in the same category *because* they are similar. We agree with the first part of this intuition but not the second. In fact, in this section and the following one we attempt to forge a rapprochement between explanation-based and similarity-based categorization. We will discuss a variety of phenomena such as representativeness, mental simulation, homeopathy and contagion, remindings, and the transfer of knowledge between domains, that reveal interactions between similarity and

explanation. In each of these cases similarity will be observed to intrude into explanatory systems. However, in most instances this will be a different type of similarity. It will be a similarity that is primarily characterized by interrelated features that are generated by underlying explanations and theories rather than by sets of independent features.

A. Similarity and decision-making

There is a ubiquitous tendency for people to focus on the resemblances or similarities between two events in order to make predictions about class membership. This reliance on similarity has a direct impact on the formation of causal explanations. When two events are similar to each other there is a strong tendency to assume that the events are associated, and this association is often assumed to be a causal one (especially in the case where one of the events precedes the other). It is very easy to move from an observation of similarity to a conclusion of causality.

Tversky and Kahneman (1973) refer to the tendency for similarity to influence people's informal decision-making and judgments as the representativeness heuristic. This use of similarity may fly in the face of logic. For example, people judge that it is more likely that a person is both over 55 and has a heart attack than the simple event of having a heart attack. This suggests that likelihood judgments are based on similarity to some prototypic victim of a heart attack.

As a further illustration consider the following example (from Kahneman and Tversky, 1982). Two people share a cab to the airport where they are to take separate flights which happen both to be scheduled for an 11:00 a.m. departure. Traffic is unusually heavy and the associated delays cause them to arrive at the airport 45 minutes late. One person finds out that his plane left on time and the other learns that his plane was delayed and departed just five minutes earlier. Which person is likely to be more unhappy about missing his plane? Other things being equal, most people would agree that it would be the person whose plane had been delayed. Kahneman and Miller (1986) interpret this example in terms of mental simulation and possible worlds. It is easier for the person who just missed his plane to imagine similar possible worlds in which he would have been able to catch the plane than it is for the other person. Again, similarity seems to intrude on reasoning and causal explanation. Note however that the similarity which enters this scenario generation and evaluation of possible worlds is a more complicated type of similarity.

B. Homeopathy and contagion

The impact of more direct perceptual similarity on the development of causal explanations is evident in the structure of people's naive theories. For example, in *The New Golden Bough*, Frazer's (1959) analysis of cultural belief systems led to the principle of homeopathy (cause and effect tend to be similar). One manifestation of this principle is homeopathic medicine. For example, in the Azande culture the cure for epilepsy is to eat the ashes of the burnt skull of the red bush monkey. Why? Because the red bush monkey happens to exhibit seizure-like stretching motions in the morning. Thus, the cure (and often the cause) is seen to resemble the symptoms. Similarly, the Azande cure for ringworm is to apply fowl's excrement (the ringworm looks like the excrement). There are numerous other examples of homeopathic medicine, such as in curing jaundice with yellow substances, stopping bleeding through the use of red stones, and curing skin growths by rubbing across the growth with a downward motion at the instant a shooting star is viewed.

Similarity also affects explanations in other ways. For example, people often make the implicit assumption that there should be similarity between the magnitudes of effects and causes. If the effect is large the cause should be large, or if the effect is complex the cause should be complex. Thus, as pointed out by Einhorn and Hogarth (1986), when germ theory was first introduced it met with a great deal of resistance and disbelief due to the discrepancy in the magnitude of the causes and effects. It was inconceivable that tiny invisible organisms (a small cause) could have such powerful and devastating effects as illness and death (a large effect).

Another set of factors that influence the form of causal explanations was referred to by Frazer as contagion (a cause must have some form of contact to transmit its effect). Thus once an arrow pierced a man it was believed that the fate of the man was held by the arrow. If the arrow was subsequently damaged, for example, then the wounded person would be caused great suffering. Similarly, in many cultures items that have had physical contact with a person (e.g., clothes, strands of hair, and teeth) were believed to assume the characteristics of the person. Thus the Papuans of Tumeleo search

desperately for any scrap of their clothing that may have been caught on a branch, for if an enemy were to find this scrap the person would be at the mercy of the enemy. In general, the more contiguous events are in time and space, the more likely it is that they will be perceived as causally related (e.g., Michotte, 1963; see also Shanks and Dickinson, this volume). If these examples seem exotic, see Rozin, Millman and Nemeroff (1986) for evidence that contagion operates in contemporary Western cultures as well.

It should be noted that whilst factors such as similarity, temporal order, and contiguity influence the development of causal explanations, their influence is constrained by broader knowledge. For example, if one has contact with a person afflicted by a cold and starts to sneeze thirty minutes later, biological knowledge prevents a causal attribution despite the influence of factors such as temporal order, contiguity and similarity.

Shweder (1977) maintains that resemblance, not co-occurrence, is the fundamental conceptual tool of everyday thinking in all cultures, not just so-called primitive cultures. That is, instead of relying on empirical observations of co-occurrences, people depend on conceptual similarities or resemblance in forming beliefs. Thus, commonsense notions like permissive child-rearing practices lead to anarchy or liberal political views, and that people who are leaders will have high self-esteem, are based on conceptual similarities rather than observations of co-occurrences. Whilst some of these characteristics might in fact covary, the basis for their purported relationship is conceptual similarity rather than scientific explanation. Thus in Shweder's view the Azandes' perceived connection between ringworm and fowl's excrement is much like our perceived connection between self-esteem and leadership in that both are based on resemblances rather than on the observation of co-ocurrence.

c. Similarity and access to explanations

Similarity is likely to have a significant effect on explanations in another way. Given the importance of similarity relations in retrieval (where similarity is defined in a theory-based way that allows for the encoding of relational properties), it is likely that explanations that are applied to a novel event are likely to be constrained by similar events and their associated explanations. Thus, if a characteristic of a person *Y* reminds you

of a previously known person *X* and an attribution had been made to explain this characteristic of person *X*, then the same attribution is likely to be applied to person *Y*. In support of the importance of reasoning from prior episodes, Read (1983) found that in some cases people rely on single, similar instances in making causal attributions.

One important question in this mode of transferring explanations is what principles govern access to prior episodes. Is access determined by superficial similarities or more abstract, perhaps causal, relations? Gentner and Landers (1985) examined this issue in a recent study and found that the degree to which a story cued recall of a previously read scenario was more dependent on literal or surface similarity rather than on higher-order similarities defined in terms of causal relations (see also Ross, 1984). Judged aptness of analogy depended on these higher-order relations. Similarity-based access to explanations may lead to the application of incorrect explanations to situations, or at least constrain the range of explanations considered.

This line of work highlights the transfer of knowledge from domain to domain based on either superficial similarity or higher-order similarity based on causal relations. Consistent with this idea, analogies can be viewed as having the ability to influence the form of theories. Consider, for example, theories about extroversion and introversion. In terms of a reservoir metaphor, extroverts are viewed as having excess energy that is expressed socially. If one takes a homeostatic system as a metaphor, however, then extroverts would be viewed as being chronically under-aroused and thus seeking social stimulation. The point of this example is that theories and explanations might be constrained by the structure of knowledge in other domains, and *access* to this knowledge seems to be influenced by similarity.

By now we may have left the reader in a somewhat uneasy state. We started this paper by criticizing the notion of similarity and arguing that concepts are organized around people's theories about the world. Yet in discussions of explanation-based categorization and reasoning, similarity was observed to influence the structure of explanations. We have not, however, come full circle. The similarity intruding into explanation and the similarity to be considered in the next section are not necessarily the similarity that defined concepts in terms of independent attributes. Before

discussing this alternative notion of similarity we shall first provide a framework or rationale that links similarity and explanation.

V. INTEGRATION OF EXPLANATION AND SIMILARITY

Our main idea can be conveyed with a simple example. It is not the case that two people are twins *because* they are very similar to each other. Rather, two people may be very similar to each other *because* they are twins. That is, similarity can be seen as the product of a deeper underlying (in this case, genetic) mechanism.

Our notion is akin to the distinction between the core and identification procedure (e.g., Miller and Johnson-Laird, 1976). The core of a concept contains abstract defining features that bring out relations between a concept and other concepts. The identification procedure consists of perceptual and more characteristic features that are used to pick out instances of a concept.

It is important to be specific about our ideas here because the kinship to core versus identification procedures could be misleading. First of all, we do not take the distinction to correspond to the difference between metaphysics and epistemology. We view both surface similarity and notions about underlying mechanisms as matters of epistomology, not metaphysics. Second, there may be a tendency to view identification versus core properties as being analogous to the *combination* to a safe versus the *contents* of the safe. By this analogy the two types of properties have little to do with each other. Instead, we maintain that core and identification properties often are intimately linked and that the core properties may give rise to identification properties. For example, the concept *man* can be defined in terms of core features like "adult, male, human", and identification properties like hair length, presence of a beard or moustache, or characteristic gait may be used to decide that some person is a man. These properties are not unrelated inasmuch as being male is partly a matter of hormones that influence physical attributes such as facial hair, and being male in our Western culture at least partly constrains other properties such as hair length and characteristic gait.

We are not arguing that the similarity view is after all correct.

Appearances can be deceiving, and members of the same category (e.g., bean bag chairs and kitchen chairs) need not necessarily look alike. Our claim is that similarity is more a good heuristic than an iron-clad guide to categorization.

The view we are espousing represents a form of psychological *essentialism*. As used by Linnaeus, the essence of a concept is its "real nature" or that which makes the thing what it is. In Aristotelean logic the essence gives rise to properties that are inevitable consequences. We are going to have to use the notion of essence a little more loosely because we want to extend our scope beyond natural kinds to artifacts such as chairs. For artifacts the basic idea is that the intentions of the builder, the goals of the user, and the functions which an object must satisfy serve to structure and give definition to the concept and help constrain more superficial or surface properties.

Several implications seem to follow fairly naturally from this form of functional essentialism and the notion of conceptual cores. First of all, this view is antithetical to independent attributes. Cuvier, a contemporary of Linnaeus, argued that every organized being forms a whole and that each part taken separately indicates and implies all the others. By this rationale, family-resemblance categories may be organized not so much in terms of surface features but in terms of underlying principles. For example, characteristic properties of the category *bird*, such as having hollow bones, feathers, wings, building nests in trees, and even singing can all be seen as adaptations to allow for or respond to consequences of flying. Our studies which finally succeeded in obtaining family resemblance sorting can be interpreted as an indication that the apparent use of family resemblance rules may be masking the use of a deeper principle that some core factor or cause is present which probabilistically leads to surface structure (family resemblance features).

Once the notion of independent properties is abandoned it becomes natural to think in terms of relational properties. Gentner (1983) has argued that analogies and metaphors are organized around relational properties rather than attributes. As an illustration of this idea, consider the Rutherford analogy between the atom and the solar system. In Gentner's formulation, features that belong to a system of mutually interconnected causal relations are likely to be transferred from one domain to another. For

example, properties such as "revolves around" (planets, sun), "attractive force" (sun), "distance" (sun, planets) and "more massive than" (sun, planets) are all causally connected based on our understanding of central force systems. Thus these causal relations will be transferred to our understanding of an electron, and properties such as yellow and hot will be ignored. Thus Gentner's model defines similarity in terms of patterns of causal relations between features rather than in terms of independent features. Given that theories lead to the development of inter-property linkages and that the encoding of objects is generated by our theories, it seems that it is important to define similarity in terms of the relational properties in which features participate. There is no compelling reason why similarity notions must be linked with the practice of treating features as independent.

Just as Gentner has argued that higher-order relations are more important than lower-order relations, so also is it natural to argue from our view of psychological essentialism that some properties are more central to a concept than others. For example, Asch and Zukier (1984) presented subjects with discordant or apparently antagonistic descriptions of a person (e.g., generous, vindictive) and asked the subjects whether the descriptions made sense. These discrepancies were nearly always readily resolved and one of the consequences of these modes of resolution was that one attribute might emerge as more important than the other. For example, when the attributes were *generous* and *vindictive*, a typical resolution was to focus on the vindictiveness as central and to assume that generosity only derived from ulterior motive. Intuition also suggests that a property that is equally true of two different concepts may be more central to one concept than to the other. For example, to our knowledge all *bananas* are curved and all (conventional) *boomerangs* are curved, but a *straight banana* would continue to satisfy the notion of a banana in a way that a *straight boomerang* would not.

These observations suggest that by linking similarity to explanations, goals, and causal structures we can arrive at just those kinds of similarity that will prove helpful in understanding how people learn about and use concepts. Perhaps this notion of similarity makes intuitive sense, but to avoid logical nonsense, however, we shall need to go a step further. So far we have argued that more central properties or explanatory principles may give rise to superficial similarity, but we still need to have organisms

sensitive to the right kinds of similarity. Otherwise the theorem of the Ugly Duckling, and Nelson Goodman's strictures on similarity, will still leave us in a quandary.

A. Tuning of similarity

What we have said so far suggests something about the kinds of similarity that it might be useful to attend to, namely those that derive from underlying cores, essences or causal principles. An organism tuned to the relevant forms of similarity would be able to store information that preserved explanatory principles. If humans are somehow attuned to the right kinds of similarity (through the nature of our perceptual and conceptual systems and through the influence of one's culture and environment), then there would be close links between similarity and explanation. (See Lebowitz, 1986a, b, for related arguments from an artificial intelligence perspective.) Appropriately constrained similarity would allow a novice to set up useful categories even in the absence of knowledge of underlying principles.

Recent research in developmental psychology is highly consistent with this general claim. Infants and young children are novices in many domains and being sensitive to appropriate types of similarity may help them to construct just those categories that will be useful later on. What kind of similarity do novices need? If young children were, in general, highly selective and focussed on a single stimulus dimension, then they might attend to the wrong dimensions and miss the relevant ones. Furthermore, although some patterns of correlated attributes are no doubt important, Keil (1981) points out that there are many possible correlations and no guarantees that children would find the appropriate ones. A surprisingly powerful strategy (pointed out and argued for by Brooks, 1978, and Kemler-Nelson, 1984) is to respond in terms of overall similarity. For example, Medin (1983) demonstrated that storing instances and accessing them in terms of one form of overall similarity (interactive rather than independent) leads to sensitivity to correlated attributes in the absence of any explicit analysis of correlation or co-occurrence.

In support of this general argument, it has often been argued that as children get older they exhibit a developmental shift from holistic representations based on overall similarity to analytic representations (e.g., Kemler and Smith, 1978; Smith and Kemler-

Nelson, 1984; Vygotsky, 1962). Holistic representations preserve characteristic features along several dimensions rather than relying on a single defining feature. In support of this idea Kemler-Nelson (1984) found that five-year-olds had more difficulty learning categories defined by a single feature than categories defined by overall similarity, but this was not true for ten-year-olds. In some related research, Keil and Batterman (1984) have found evidence for a developmental shift in the representation of word meanings from collections of characteristic features to more nearly defining features. For example, kindergartners (5-6 years old) preferred a description of an island as a place that is warm, and has coconut trees, palm trees, and girls with flowers in their hair, even though the technical definition of an island was violated in the description. In contrast, fourth graders (9-10 years old) preferred a description that had none of the characteristic features of an island, but contained the crucial information that the land was surrounded by water on all sides. There is also evidence that infants respond in terms of overall similarity (e.g., Bornstein, 1984; Offenbach and Blumberg, 1985; Younger and Cohen, 1985) even before they have an explicit notion of what a stimulus dimension might be (Smith, 1986). Furthermore, children find classification tasks involving superordinate categories, where overall similarity is not such a good guide, to be more difficult than tasks involving basic level categories (Horton and Markman, 1980; Rosch *et al.*, 1976).

Evidence with adult subjects is also consistent with these ideas. Wattenmaker and Medin (1987) found that if adults learn about examples through a procedure that encourages responding on the basis of overall similarity, then family-resemblance constructions can be obtained even in the absence of conceptual or theory-driven underpinnings. Specifically, in this paradigm (borrowed from Brooks, 1978) subjects initially learned to associate a label with each of the examples (short descriptions of hypothetical people) and when this paired-associate task was mastered, participants performed the sorting task with only the associated labels and not the original examples. This initial learning forced subjects to process the individual examples as bounded entities rather than dissecting the examples into their components. In a study using similar procedures Wattenmaker (1987) found that complex correlations between features influenced categorization even when the correlation was not selectively encoded and could not be verbalized. Both of these results suggest that if subjects are attuned to the right kinds of similarity, holistic representations based on overall similarity can preserve useful categories even when the underlying principles are not understood. In many cases, we have argued, there will be deeper principles linked to or generating superficial similarity. Therefore organisms that are biased or predisposed to form concepts on the basis of overall similarity will have an appropriate parsing for a later stage when greater knowledge will allow a deeper basis for categorization.

The one exception to this general claim concerning the importance of overall similarity indirectly serves to support it. Overall similarity is a good general-purpose heuristic but certain relationships in our environment have presumably been stable for many thousands of years. Examples might be the distinction between animate and inanimate objects and certain principles of physical causality such as the fact that solid objects do not move through the space occupied by other solid objects. These forms of stability ought to permit perceptual analyses that are either innate or acquired very early on. Indeed there is evidence that such selectivity (and not just overall similarity) for these types of information is present in very young infants (e.g., Baillargeon, Spelke and Wasserman, 1985; Gelman and Spelke, 1981; Gibson and Spelke, 1983; and Leslie, 1982).

VI. CONCLUSIONS

In this chapter we have argued that a similarity-based approach to conceptual coherence is insufficient to explain the richness of conceptual structure, and opted instead for a knowledge-based approach to conceptual coherence. A knowledge-based approach emphasizes that coherence derives from both the internal causal structure of a conceptual domain and the position of the concept in the complete knowledge base. Concepts are viewed as embedded in theories, and are coherent to the extent that they fit people's background knowledge or naive theories about the world.

In discussing theory-based categorization, however, similarity was often observed to constrain and structure explanations. We were therefore led to explore formulations of similarity that are consistent with an emphasis on flexibility and theory-driven

categorization. Our proposed rapprochement between similarity and explanations is that explanations form part of the deep structure which defines deep similarity (relational structures) and may generate surface similarity. This modified notion of similarity may constrain explanation by influencing patterns of association (especially causal associations) and by constraining access to prior explanations. The notion of an underlying explanatory structure seems to direct us toward just those forms of similarity that are likely to be useful in conceptual functioning.

Where does all this exposition leave us with respect to the question of conceptual coherence? We believe that a complete account of conceptual coherence will require an understanding of a similarity-based component and a knowledge-based component. Superficial similarity is not an infallible guide to categorization but it is right enough often enough to provide a useful heuristic. Even so, we may be guilty of overstating our case in that whilst surface similarity probably has a role to play in the categorization of natural kinds and artifacts, it may be of little use in structuring concepts such as verbs (see Gentner, 1981). In any event, even perceptual similarity seems to require relational properties and not just attributes. Inter-property relationships are important ingredients at the level of similarities and rules as well as at the level of knowledge or theories.

Our preferred reading of our thesis is not that there is a pendulum between similarity- and explanation-based categorization that swings one way or another depending on whim or fancy, but rather that we are slowly tracing our way up a spiral with ever more accurate conceptions of similarity and explanation.

REFERENCES

Armstrong, S.L., Gleitman, L.R. and Gleitman, H. (1983). "What some concepts might not be". *Cognition, 13,* 263–308.

Asch, S.E. and Zukier, H. (1984). "Thinking about persons". *Journal of Personality and Social Pychology, 45,* 1230–40.

Baillargeon, R., Spelke, E.S. and Wasserman, S. (1985). "Object permanence in five-month-old infants". *Cognition, 20,* 191–208.

Barresi, J. (1981). "Perception and imagination". Paper presented at the Conference on the Philosophy of Perception and Psychology. Montreal, Canada.

Barsalou, L.W. (1983). "Ad hoc categories". *Memory and Cognition, 11,* 211–27.

Barsalou, L.W. (1985). "Ideals, central tendency, and frequency of instantiation as determinants of graded structure in categories". *Journal of Experimental Psychology: Learning, Memory, and Cognition, 11,* 629–54.

Bornstein, M.H. (1984). "A descriptive taxonomy of psychological categories used by infants". In C. Sophian (Ed.), *Origins of Cognitive Skills.* (pp. 313–38). Hillsdale, N.J.: Lawrence Erlbaum.

Brooks, L. (1978). "Nonanalytic concept formation and memory for instances". In E. Rosch and B.B. Lloyd (Eds), *Cognition and Categorization.* (pp. 169–215). Hillsdale, N.J.: Lawrence Erlbaum.

Chapman, L.J. and Chapman, J.P. (1967). "Genesis of popular but erroneous psychodiagnostic observations". *Journal of Abnormal Psychology, 72,* 193–204.

Chapman, L.J. and Chapman, J.P. (1969). "Illusory correlation as an obstacle to the use of valid psychodiagnostic signs". *Journal of Abnormal Psychology, 74,* 271–80.

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

Frazer, J.G. (1959). *The New Golden Bough.* New York: Criterion Books.

Gati, I. and Tversky, A. (1984). "Weighting common and distinctive features in perceptual and conceptual judgments". *Cognitive Psychology, 16,* 341–70.

Gelman, R. and Spelke, E. (1981). "The development of thoughts about animate and inanimate objects: Implications for research on social cognition". In J.H. Flavell and L. Ross (Eds), *Social Cognitive Development.* (pp. 43–66). New York: Cambridge University Press.

Gentner, D. (1981). "Some interesting differences between verbs and nouns". *Cognition and Brain Theory, 4,* 161–78.

Gentner, D. (1983). "Structure-mapping: A theoretical framework for analogy". *Cognitive Science, 7,* 155–70.

Gentner, D. and Landers, R. (1985). "Analogical reminding: A good match is hard to find". Paper presented at the International Conference of Systems, Man and Cybernetics. Tucson, Arizona.

Gibson, E.J. and Spelke, E.S. (1983). "The development of perception". In J.H. Flavell and E.M. Markman (Eds), *Cognitive Development.* (Vol. III, pp. 1–76). P.H. Mussen (Series Ed.), *Handbook of Child Psychology.* New York: Wiley.

Horton, M.S. and Markman, E.M. (1980). "Developmental differences in the acquisition of basic and superordinate categories". *Child Development, 51,* 708–19.

Johnson-Laird, P.N., Herrmann, D.J. and Chaffin, R. (1984). "Only connections: A critique of semantic networks". *Psychological Bulletin, 96,* 292–315.

Kahneman, D. and Miller, D.T. (1986). "Norm theory: Comparing reality to its alternatives". *Psychological Review, 93,* 136–53.

Kahneman, D. and Tversky, A. (1982). "The simulation heuristic". In D. Kahneman, P. Slovic and A. Tversky (Eds), *Judgment under Uncertainty: Heuristics and Biases.* (pp. 201–08). Cambridge: Cambridge University Press.

Kant, I. (1788/1933). *Critique of Pure Reason.* Trans. Norman Kemp Smith. London: Macmillan.

Katz, J.J. and Postal, P.M. (1964). *An Integrated Theory of Linguistic Descriptions.* Cambridge, MA: MIT Press.

Keil, F.C. (1981). "Constraints on knowledge and cognitive development". *Psychological Review, 88,* 197–227.

Keil, F.C. and Batteiman, N. (1984). "A characteristic-to-defining shift in the development of word meaning". *Journal of Verbal Learning and Verbal Behavior, 23,* 221–36.

Kemler, D.G. and Smith, L.B. (1978). "Is there a developmental trend from integrality to separability in perception?". *Journal of Experimental Child Psychology, 26,* 460–507.

Kemler-Nelson, D.G. (1984). "The effect of intention on what concepts are acquired". *Journal of Verbal Learning and Verbal Behavior, 23,* 734–59.

Lebowitz, M. (1986a). "Not the path to perdition: The utility of similarity-based learning". Manuscript submitted for publication.

Lebowitz, M. (1986b). "Integrated learning: Controlling explanation". *Cognitive Science, 10,* 219–40.

Leslie, A.M. (1982). "The perception of causality in infants". *Perception, 11,* 173–86.

Medin, D.L. (1983). "Structural principles in categorization". In T. Tighe and B.E. Shepp (Eds), *Perception, Cognition, and Development: Interactional Analyses.* (pp. 203–30). Hillsdale, N.J.: Lawrence Erlbaum.

Medin, D.L. and Schwanenflugel, P.J. (1981). "Linear separability in classification learning". *Journal of Experimental Psychology: Human Learning and Memory, 7,* 355–68.

Medin, D.L. and Smith, E.E. (1984). "Concepts and concept formation". *Annual Review of Psychology, 35,* 113–38.

Medin, D.L. and Wattenmaker, W.D. (1987). "Category cohesiveness, theories and cognitive archeology". In U. Neisser (Ed.), *Concepts Reconsidered: The Ecological and Intellectual Bases of Categories.* Cambridge: Cambridge University Press.

Medin, D.L., Wattenmaker, W.D. and Hampson, S. (1987). "Family resemblance, concept cohesiveness, and category construction". *Cognitive Psychology, 19,* 242–279.

Mervis, C.B. and Rosch, E. (1981). "Categorization of natural objects". *Annual Review of Psychology, 32,* 89–115.

Michotte, A.E. (1963). *The Perception of Causality.* New York: Basic Books.

Miller, G.A. and Johnson-Laird, P.N. (1976). *Language and Perception.* Cambridge, MA: Harvard University Press.

Murphy, G.L. (1982). "Cue validity and levels of categorization". *Psychological Bulletin, 91,* 174–7.

Murphy, G.L. and Medin, D.L. (1985). "The role of theories in conceptual coherence". *Psychological Review, 92,* 289–316.

Nakamura, G.V. (1985). "Knowledge-based classification of ill-defined categories". *Memory and Cognition, 13,* 377–84.

Nilsson, N.J. (1965). *Learning Machines.* New York: McGraw-Hill.

Offenbach, S.I. and Blumberg. F.C. (1985). "The concept of dimensions in developmental research". In H.W. Reese (Ed.), *Advances in Child Development and Behavior.* (Vol. XIX, pp. 83–112). New York: Academic Press.

Ortony, A., Vondruska, R.J., Foss, M.A. and Jones, L.E. (1985). "Salience, similes, and the asymmetry of similarity". *Journal of Memory and Language, 24,* 569–94.

Read, S.J. (1983). "Once is enough: Causal reasoning from a single instance". *Journal of Personality and Social Psychology, 45,* 323–34.

Rips, L.G. and Handte, J. (1984). "Classification without similarity". University of Chicago. Unpublished manuscript.

Rosch, E. (1975). "Cognitive representations of semantic categories". *Journal of Experimental Psychology: General, 104,* 192–233.

Rosch, E. (1978). "Principles of categorization". In E. Rosch and B.B. Lloyd (Eds), *Cognition and Categorization.* (pp. 27–48). Hillsdale, N.J.: Lawrence Erlbaum.

Rosch, E. and Mervis, C.B. (1975). "Family resemblances: Studies in the internal structure of categories". *Cognitive Psychology, 7,* 573–605.

Rosch, E., Mervis, C.B., Gray, W.D., Johnson, D.M. and Boyes-Braem, P. (1976). "Basic objects in natural categories". *Cognitive Psychology, 8,* 382–439.

Ross, B.H. (1984). "Remindings and their effects in learning a cognitive skill". *Cognitive Psychology, 16,* 37–416.

Roth, E.M. and Shoben, E.J. (1983). "The effect of context on the structure of categories". *Cognitive Psychology, 15,* 346–78.

Rozin, P., Millman, L. and Nemeroff, C. (1986). "Operation of the laws of sympathetic magic in disgust and other domains". *Journal of Personality and Social Psychology, 50,* 703–12.

categorization. Our proposed rapprochement between similarity and explanations is that explanations form part of the deep structure which defines deep similarity (relational structures) and may generate surface similarity. This modified notion of similarity may constrain explanation by influencing patterns of association (especially causal associations) and by constraining access to prior explanations. The notion of an underlying explanatory structure seems to direct us toward just those forms of similarity that are likely to be useful in conceptual functioning.

Where does all this exposition leave us with respect to the question of conceptual coherence? We believe that a complete account of conceptual coherence will require an understanding of a similarity-based component and a knowledge-based component. Superficial similarity is not an infallible guide to categorization but it is right enough often enough to provide a useful heuristic. Even so, we may be guilty of overstating our case in that whilst surface similarity probably has a role to play in the categorization of natural kinds and artifacts, it may be of little use in structuring concepts such as verbs (see Gentner, 1981). In any event, even perceptual similarity seems to require relational properties and not just attributes. Inter-property relationships are important ingredients at the level of similarities and rules as well as at the level of knowledge or theories.

Our preferred reading of our thesis is not that there is a pendulum between similarity- and explanation-based categorization that swings one way or another depending on whim or fancy, but rather that we are slowly tracing our way up a spiral with ever more accurate conceptions of similarity and explanation.

REFERENCES

Armstrong, S.L., Gleitman, L.R. and Gleitman, H. (1983). "What some concepts might not be". *Cognition, 13,* 263–308.

Asch, S.E. and Zukier, H. (1984). "Thinking about persons". *Journal of Personality and Social Psychology, 45,* 1230–40.

Baillargeon, R., Spelke, E.S. and Wasserman, S. (1985). "Object permanence in five-month-old infants". *Cognition, 20,* 191–208.

Barresi, J. (1981). "Perception and imagination". Paper presented at the Conference on the Philosophy of Perception and Psychology. Montreal, Canada.

Barsalou, L.W. (1983). "Ad hoc categories". *Memory and Cognition, 11,* 211–27.

Barsalou, L.W. (1985). "Ideals, central tendency, and frequency of instantiation as determinants of graded structure in categories". *Journal of Experimental Psychology: Learning, Memory, and Cognition, 11,* 629–54.

Bornstein, M.H. (1984). "A descriptive taxonomy of psychological categories used by infants". In C. Sophian (Ed.), *Origins of Cognitive Skills.* (pp. 313–38). Hillsdale, N.J.: Lawrence Erlbaum.

Brooks, L. (1978). "Nonanalytic concept formation and memory for instances". In E. Rosch and B.B. Lloyd (Eds), *Cognition and Categorization.* (pp. 169–215). Hillsdale, N.J.: Lawrence Erlbaum.

Chapman, L.J. and Chapman, J.P. (1967). "Genesis of popular but erroneous psychodiagnostic observations". *Journal of Abnormal Psychology, 72,* 193–204.

Chapman, L.J. and Chapman, J.P. (1969). "Illusory correlation as an obstacle to the use of valid psychodiagnostic signs". *Journal of Abnormal Psychology, 74,* 271–80.

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

Frazer, J.G. (1959). *The New Golden Bough.* New York: Criterion Books.

Gati, I. and Tversky, A. (1984). "Weighting common and distinctive features in perceptual and conceptual judgments". *Cognitive Psychology, 16,* 341–70.

Gelman, R. and Spelke, E. (1981). "The development of thoughts about animate and inanimate objects: Implications for research on social cognition". In J.H. Flavell and L. Ross (Eds), *Social Cognitive Development.* (pp. 43–66). New York: Cambridge University Press.

Gentner, D. (1981). "Some interesting differences between verbs and nouns". *Cognition and Brain Theory, 4,* 161–78.

Gentner, D. (1983). "Structure-mapping: A theoretical framework for analogy". *Cognitive Science, 7,* 155–70.

Gentner, D. and Landers, R. (1985). "Analogical reminding: A good match is hard to find". Paper presented at the International Conference of Systems, Man and Cybernetics. Tucson, Arizona.

Gibson, E.J. and Spelke, E.S. (1983). "The development of perception". In J.H. Flavell and E.M. Markman (Eds), *Cognitive Development.* (Vol. III, pp. 1–76). P.H. Mussen (Series Ed.), *Handbook of Child Psychology.* New York: Wiley.

Horton, M.S. and Markman, E.M. (1980). "Developmental differences in the acquisition of basic and superordinate categories". *Child Development, 51,* 708–19.

Sebestyn, G.S. (1962). *Decision-making Processes in Pattern Recognition.* New York: Macmillan.

Shweder, R.A. (1977). "Likeness and likelihood in everyday thought: Magical thinking in judgments about personality". *Current Anthropology, 18,* 4.

Smith, E.E. and Medin, D.L., (1981). *Categories and Concepts.* Cambridge, MA: Harvard University Press.

Smith, J.D. and Kemler-Nelson, D.G. (1984). "Overall similarity in adults' classification: The child in all of us". *Journal of Experimental Psychology: General, 113,* 137–59.

Smith, L.B. (1986). "From global similarities to kinds of similarities: The construction of dimensions". Paper presented at the Workshop on Similarity and Analogy, University of Illinois (June).

Tversky, A. (1977). "Features of similarity". *Psychological Review, 84,* 327–52.

Tversky, A. and Kahneman, D. (1973). "Availability: A heuristic for judging frequency and probability". *Cognitive Psychology, 5,* 207–32.

Tversky, B. and Hemenway, K. (1984). "Objects, parts, and categories". *Journal of Experimental Psychology: General, 113,* 169–93.

Vygotsky, L.S. (1962). *Thought and Language.* Cambridge, MA: MIT Press.

Watanabe, S. (1969). *Knowing and Guessing: A Formal and Quantitative Study.* New York: Wiley.

Wattenmaker, W.D. (1987). "Nonanalytic concept formation and sensitivity to correlated attributes". Manuscript in preparation. University of Illinois.

Wattenmaker, W.D., Dewey, G.I., Murphy, T.D. and Medin, D.L. (1986). "Linear separability and concept learning: Context, relational properties, and concept naturalness". *Cognitive Psychology, 18,* 158–94.

Wattenmaker, W.D. and Medin, D.L. (1987). "Family resemblance and nonanalytic category construction". Manuscript in preparation. University of Illinois.

Wittgenstein, L. (1953). *Philosophical Investigations.* G.E.M. Anscombe. Oxford: Basil Blackwell.

Younger, B.A. and Cohen, L.B. (1985). "How infants form categories". In G.H. Bower (Ed.), *The Psychology of Learning and Motivation.* (Vol. XIX, pp. 211–47). New York: Academic Press.