Studies in Cognition

Editor: Glyn Humphreys
University of Birmingham

Knowledge, concepts and categories Koen Lamberts & David Shanks (eds)
Cognitive models of memory Martin A. Conway (ed.)
Cognitive neuroscience Michael D. Rugg (ed.)
Attention Harold Pashler (ed.)

Knowledge, Concepts and Categories

Edited by

Koen Lamberts
University of Birmingham

David Shanks
University College London

Psychology Press
a member of the Taylor & Francis group

CHAPTER TWO

Concepts and Similarity

Ulrike Hahn and Nick Chater

CONCEPTS AND SIMILARITY – THE CHICKEN AND THE EGG?

The cognitive system does not treat each new object or occurrence as distinct from and unrelated to what it has seen before: it classifies new objects in terms of concepts which group the new object together with others which have previously been encountered. Moreover, the cognitive system also judges whether new objects are similar to old objects. *Prima facie*, these processes seem to be related, but exactly how they are related is not so clear. This puzzle is important because concepts are thought to be building blocks in terms of which knowledge is represented (e.g. Oden 1987).

One suggestion concerning the relationship between concepts and similarity is that concepts group together objects which are similar. According to this point of view, the reason that "bird" is a useful concept is that birds are relatively similar to each other – mostly having wings, laying eggs, building nests, flying and so on. A hypothetical concept "drib" which grouped together a particular lightbulb, Polly the pet parrot, the English channel and the ozone layer would seem to be a useless, and highly bizarre, concept precisely because the items it groups together are not at all similar.[1] Why is it important that concepts group together similar things – why is "bird" a more coherent concept than "drib"? One suggestion is that birds, being similar, support many interesting generalizations (most birds have wings, most birds fly, and so on); but there seem to be no interesting generalizations to state about dribs. Moreover, on learning that Polly has a beak, it is reasonable to infer that other birds also may have beaks (since Polly is a bird, and birds are similar to each other); on the other hand, it is not reasonable to infer that other dribs may have beaks, since other dribs and Polly have nothing in common. If this view of the relation between concepts and similarity is correct, then similarity is at the very centre of the theory of concepts: a theory of similarity would explain, or at least be an important factor in explaining, why people have the concepts that they do.

There is, however, an alternative view of the relationship between concepts and similarity that also has considerable intuitive appeal. What is it for two objects to be similar? Presumably it is that they have many properties in common – indeed, this point of view is implicit in our discussion above. But to say that birds are similar because, among other things, birds generally lay eggs is the same as saying that birds are similar because, among other things, they are grouped together by the concept "egg-layer". In the same way, the similarity of birds seems to be rooted in the fact that most birds are members of concepts "flyer, has wings", and so on. Thus, it seems that objects are similar because they fall under the same concepts.

Bringing together these intuitively plausible views confronts us with a "chicken and egg" problem. The first point of view suggests that similarity can be used to explain concepts; the second point of view suggests that concepts can be used to explain similarity. This seems dangerously circular, to say the least! The relationship between similarity and concepts is plainly not a straightforward one. As we have seen, it is not even clear which notion should be taken as fundamental. Moreover, unravelling the relationship between concepts and similarity is not merely an entertaining puzzle; it goes to the core of current theories of concepts.

Given this tight connection, it is surprising that there is little research directly integrating the two. Similarity, although frequently employed as an explanatory notion in the concepts literature, is seldom given closer scrutiny. Likewise, models of similarity typically assume given properties, and thus concepts, as a starting point. In both cases, this leaves out the question of whether or not the notion one is building on can actually fulfil its designated role. In this chapter we consider in what ways concepts and similarity are related, and how research in both areas can be brought together. This task is complicated by the fact that just as there are a range of competing theories of concepts, there are also a range of competing views of similarity, none of which is entirely satisfactory. Furthermore, different views of concepts have different roles for similarity, and not all notions of similarity are consistent with all views of concepts.

We begin by establishing the precise role attributed to similarity in current theories of conceptual structure. Subsequently, we investigate the notion of similarity, examining both general issues and current models of similarity. These will be drawn not only from psychology but also from artificial intelligence and computer science. Specifically, we introduce *neural networks*, *case-based reasoning* (CBR) and a relatively little-known account based on a mathematical notion called *Kolmogorov complexity*. These models are assessed for their adequacy as models of similarity, as only a theory of similarity which is in itself satisfactory can ground a theory of conceptual structure. We then bring these models together with the theories of conceptual structure introduced, examining which models of similarity are compatible with which views of conceptual structure. Finally, leaving particular models and theories behind, we return again to the question of the general relationship between concepts and similarity. We review the empirical evidence and establish along what lines the problem of the "chicken and the egg" might one day be resolved.

CONCEPTS

We begin by giving a brief overview of present theories of concepts in relation to similarity. Current theories of concepts are covered in more detail elsewhere in this volume, and in Komatsu (1992) and Medin (1989). We first outline two theories of concepts, prototype and exemplar theories, in which similarity is directly and explicitly involved. We then consider rule-based and theory-based accounts, which do not explicitly involve similarity, but in which, we argue, similarity plays an important, although indirect role.

Prototype and exemplar views: similarity centre-stage

The common thread linking this family of views is a direct connection between concepts and similarity: categorizing an object involves judging the similarity between that object and some other object(s). Exactly what the new object is compared with, and how that comparison is carried out distinguishes prototype and exemplar views from each other, and identifies different variants of each view. In all cases though, categorization depends on similarity.

Prototype views

The prototype view[2] assumes that each category is associated with a "prototype", a stored representation of the properties that typify members of that category. The classification of new objects as, for example, birds will depend on how similar that object is to the bird prototype, and also to the prototypes of other categories.

Opinions vary concerning exactly what a prototype is.[3] The simplest view assumes that prototypes are stored mental representations of the same nature as the mental representation of specific objects. It follows from this view that judging the similarity between a specific object to a prototype in classification is exactly the same process as judging the similarity between two objects.

By contrast, some prototype views assume that prototypes are not represented in quite the same way as specific objects, but are specified in somewhat more abstract terms (e.g. Taylor 1989). In its simplest form, this abstract specification could simply list certain properties which have previously appeared in instances of the category (while other features might be ignored entirely – this is what makes the representation abstract). The various listed properties might also have different "weights", reflecting their varying degrees of relevance for category membership. For example, the concept bird might consist of the following list of weighted features:

has wings	0.8
has feathers	0.9
flies	0.5
sings	0.5
lays eggs	0.9

The categorization of a new creature, then, can be divided into three stages: first, which of these features the creature possesses is assessed. Then, using this information, the similarity between the prototype and the new creature must be calculated. Many different measures of similarity, such as addition or multiplication of the weighted features that the creature possesses, have been proposed. Finally, it is necessary to decide whether the creature ultimately is "similar enough" to the bird prototype to count as a bird. This, for example, might involve seeing whether the object is more similar to the bird prototype than it is to any other prototype; again, the possibilities are numerous.

Exemplar views

The exemplar approach also sees classification as involving judgements of similarity to stored representations (see, for example, Brooks 1978, and Medin & Schaffer 1978. For more recent variants, see Kruschke 1992, and Nosofsky 1988, and for an overview, see Komatsu 1992). Instead of judging similarity to a single prototype representing each category, the new object is compared to many stored "exemplars" (specific previously encountered instances) of the category. If the new object is more similar to exemplar birds than to exemplars of any other category, then it will be classified as a bird. According to the exemplar view, the specification of the category is implicit in its instances; no necessary and sufficient features, or even probable features, are abstracted. The concept is learned simply by storing examples of its category members (for experimental investigations on abstraction in category acquisition, see e.g. Homa et al. 1981, Whittlesea 1987, this volume, and Medin et al. 1983).

As with prototype views, there are numerous specific proposals concerning how this framework is fleshed out into a full-blown model of classification. For example, specific models vary according to whether it is the similarity to a single, best-matching exemplar that matters (see, e.g. Hintzman & Ludlam, 1980, and many CBR systems in artificial intelligence, e.g. PROTOS, Porter et al. 1990) or whether a set of exemplars – either a fixed subset or the entire set – is matched (see for discussion, e.g. Homa et al. 1981, Jones & Heit 1993). In all cases, however, similarity assessment is given a central role.

Definitional and theory-based accounts: similarity behind the scenes?

In the last subsection, we considered views of concepts in which similarity is explicitly viewed as central to explaining concepts. Now we turn to rule-based and theory-based views of concepts, which do not make direct reference to similarity. Nonetheless, as we shall see, similarity may play an important, if less visible, role in categorization in these views.

Definitional views

The definitional or "classical" view of concepts holds that concepts possess definitions specifying features necessary and *sufficient* for the concept. This definition is the summary description of the entire class used in every instance of categorization, which proceeds simply by checking for the presence of these features in the entity in question. This view is commonly supplemented by the "nesting assumption" – that a subordinate concept (e.g. "robin") contains nested within in it the defining features of the superordinate ("bird"). The crucial point, in our context, is that concepts are explained without reference to similarity. The definitional view thus requires closer scrutiny.

First and foremost, the definitional view seems inadequate as a theory when transferred from artificial concepts in controlled experiments[4] to our everyday concepts, that is to the concepts for which we typically have words. Of the difficulties faced here, the most serious one is that almost all everyday concepts appear to be indefinable (Fodor et al. 1980). It simply does not seem possible to formulate necessary and sufficient conditions for being, for example, a chair, or a window, or a smile. This is illustrated by the fact that dictionary "definitions" of almost all terms are not really definitions at all. They do not provide necessary and sufficient conditions for category membership – instead they typically do no more than provide some relevant information about category members, which may help the dictionary user identify which concept is intended.[5]

Moreover, even for those concepts which do appear to have definitions, these definitions generally hold only with respect to a range of "background assumptions". Varying these assumptions immediately produces unclear or borderline cases:

> The noun bachelor can be defined as an unmarried adult man, but the noun clearly exists as a motivated device for categorizing people only in the context of a human society in which certain expectations about marriage and marriageable age obtain. Male participants in long-term unmarried couplings would not ordinarily be described as bachelors; a boy abandoned in the jungle and grown to maturity away from contact with human society would not be called a bachelor. (Fillmore, quoted from Lakoff 1987).

Background factors, such as the social conventions concerning marriage, will, in general, hold to varying degrees. Presumably the definition of bachelor can meaningfully be applied if the background conditions are sufficiently similar to the conventions concerning marriage current in the West. This is one way in which similarity can have a "behind the scenes" role in the definitional view – similarity applies to background assumptions underlying the application of necessary and sufficient conditions, rather than being explicitly mentioned in the definition itself.

There is also another way in which similarity would enter the theory of concepts, even if the definitional view were correct. We have so far dealt with the most direct difficulty of the definitional view: that it is difficult or impossible to define almost all concepts. But there is another argument, based on "prototype effects" against the definitional view. This argument, crudely stated is: if category membership is all or none, as the definitional view suggests, why is a robin judged to be (and treated as) a more typical bird than an ostrich? Some theorists have responded by arguing that such effects are attributed to the "identification procedure" for a concept – the procedure used to identify members of that concept; the "core" of the concept, used in reasoning, is still held to consist of a definition (Miller & Johnson-Laird 1976, Osherson & Smith 1981). This two-component account opens the door to similarity – the identification procedure may, for example, be based on prototypes or exemplars, as discussed above, with their direct reliance on similarity.

We have seen that, as a theory of everyday concepts, the definitional view appears to be inadequate. More importantly, the main problems it encounters appear to implicate similarity. Most everyday concepts such as "chair", or "smile" seem to involve networks of related and thus similar instances, but without there being a single set of defining properties. In the other case, similarity of background conditions must hold for a definition to be applicable. Finally, prototype effects, for which an explanation involving similarity suggests itself, must be accounted for.

In experiments on *artificial concept* learning, on the other hand, performance can be modelled accurately by assuming that subjects learn definitional rules. Here, the core problems plaguing the definitional view for everyday concepts are generally ruled out by the design of the materials. Nevertheless, recent research has revealed "intrusions" of similarity where subjects appear to make use of a rule in these contexts as well. For one, Nosofsky et al. (1989) found that the addition of an "exemplar" component to their rule-based account considerably improved the degree to which their model fits the experimental data. Thus, even when using a rule, subjects may also be paying attention to the similarity of new instances to previous instances. More direct evidence comes from Allen & Brooks (1991), who found in many, but not all, experimental conditions that similarity to past instances affected subjects' application of a simple explicit rule, specified by the experimenter. This ongoing influence of similarity to prior episodes, Allen & Brooks argue, may be particularly frequent (because useful) in uncertain situations where rules and definitions have only heuristic value. Incidentally, a similar ongoing role of prior episodes in addition to explicit instructions has emerged in the context of problem solving (Ross 1984, 1987, 1989, Ross et al. 1990, Ross & Kennedy 1990).

In sum, then, the definitional view, while appearing to ignore similarity, actually leaves open a number of ways in which similarity may affect concepts: in determining how definitions are interpreted, as playing a role in a concepts' "identification procedure", and as an additional factor affecting how definitions or rules are applied in actual classification.

Theory-based views

Theory- or explanation-based views of concepts reject exemplar, probabilistic and definitional views and focus instead on the relationship between concepts and our knowledge of, and theories about, the world (Murphy & Medin 1985, Wattenmaker et al. 1986, Lakoff 1987, Medin & Wattenmaker 1987, Wattenmaker et al. 1988, Wisniewski & Medin 1994; see also Heit's chapter in this volume). "Theory", here, can be taken to refer to a body of knowledge that may include scientific principles, stereotypes and informal observations of past experiences (Murphy & Medin 1985, Wisniewski & Medin 1994). Most importantly, properties of objects are not independent and thus not independently assessed in categorization but are embedded within networks of inter-property relationships which organize and link them (Wattenmaker et al. 1988). Accordingly, Lakoff's (1987) "idealized cognitive models" are another expression of the same idea (Medin & Wattenmaker 1987). For example, the concept "bird" cannot be merely a collection of "bird" features such as "has wings", "has feathers", and "has a beak", but must specify how these feature are related (e.g. that the wings are covered in feathers, the beak is not). But not only such relational aspects between features, but also their causal connections can play a crucial role in categorization (Wattenmaker et al. 1988). More fundamentally still, our prior theories influence what features we perceive in the first place (Wisniewski & Medin 1994).

How do theory-based views relate to similarity? It is frequently suggested that theory-based views undermine the role of similarity in theories of concepts. But this is misleading: explanation/theory-based approaches target simplistic views of similarity assessment such as simple counting of shared perceptual features. However, explanations or theories are neither capable of, nor intended to, replace similarity in categorization. What they suggest is that similarity itself, if it is to be relevant to concepts, must be influenced by our theories of, and knowledge about, the world (Lamberts 1994, Wattenmaker et al. 1988). Thus, theory-based views demand a *better* account of similarity, rather than *no* account of similarity, in explaining concepts.

Summary

We have seen that similarity plays an important role in theories of concepts based on prototypes, exemplars, definitions and theories. We now turn to similarity in order to establish whether or not it can really fit the bill.

SIMILARITY

We will begin our investigation of similarity with a treatment of the damning criticisms which have been voiced against similarity as an explanatory notion. With these out of the way, we then turn to specific models of similarity, assessing them for strengths and weaknesses. Finally, we will conclude this section with a discussion of crucial stages of the process of similarity assessment which are outside the scope of all current models of similarity.

Is similarity explanatory: the problem of "respects"

If theories of concepts are to rely on similarity, whether directly or indirectly, then similarity must be a coherent and explanatory notion. Within philosophy, however, grave doubts about the explanatory power of similarity have been expressed (Goodman 1972). If these doubts are well-founded, then the role of similarity in current theories of concepts, and indeed the viability of those theories, must be called into question. In this subsection, we consider Goodman's critique of similarity and how it relates to the theory of concepts.

What does it mean to say that two objects a and b are similar? Intuitively, we say that objects are similar because they have many properties in common. But, as Goodman pointed out, this intuition does not take us very far, because all entities have infinite sets of properties in common (Goodman 1972). A plum and a lawnmower both share the properties of weighing less than 100 kilos (and less than 101 kilos . . . etc.). This seems to imply that all objects are similar to all others! Of course, all entities will also have infinite sets of properties that are not in common. A plum weighs less than one kilo, while a lawnmower weighs more than one kilo (and similarly for 1.1 kilos and 1.11 kilos . . . etc.). Perhaps, then, all objects are dissimilar to all others! Pursuing our intuition about what makes objects similar has led to deep trouble.

Goodman concludes that "similarity" is thus a meaningful notion only as *similar in a certain respect*. Although similarity superficially appears to be a two-place relation, it is really a three-place relation $S(a, b, r)$ – a and b are similar in respect r. Any talk of similarity between two objects must at least implicitly contain some respect in which they are similar.

But, Goodman notes, once "respects" are introduced, it seems that similarity itself has no role to play: the respects do all the work. To say that an object belongs to a category because it is similar to items of that category with respect, for instance, to the property "red" is merely to say that it belongs to the category because it is red – the notion of similarity can be removed without loss. "Similarity", so Goodman says, is a "pretender, an imposter, a quack", "it has, indeed, its place and its uses, but is more often found where it does not belong, professing powers it does not possess" (Goodman 1972: 437). In particular, Goodman's qualms suggest that similarity may not be an explanatory construct upon which a theory of concepts can rely.

These criticisms have made their way into psychology only fairly recently through authors advocating theory-based approaches to concepts (Murphy & Medin 1985, Medin & Wattenmaker 1987), sparking what has been viewed as the "decline of similarity" within the study of concepts and categorization (Neisser 1987).

Two recent papers (Goldstone 1994a, Medin et al. 1993) have, however, re-evaluated whether Goodman's criticisms really undermine similarity *for psychology*. Perhaps a psychological notion of similarity may not be subject to the points that Goodman raises for similarity in the abstract. Two questions are particularly important. First, are there psychological restrictions on respects, or can simply anything be a respect? Medin et al. (1993) have argued that although similarity is highly flexible as a result of goals, purpose, or context, because respects are by no means fixed, this does not imply that they vary in arbitrary ways. Rather, there is a great deal of systematicity in the variation exhibited with constraints arising both from knowledge and purpose as from the comparison process itself, as we shall discuss in more detail below. Secondly, granting Goodman's claim that similarity involves respects, does this really imply that respects do all of the work, leaving no role for similarity? Goldstone (1994a) argues that people do not usually compare objects only in a single respect such as "size" but along multiple dimensions such as size, colour, shape, etc. Given multiple respects (or, alternatively, a complex respect, such as "colour", or "appearance") the psychologically central issue is how different factors are combined to give a single similarity judgement (Goldstone 1994a). Thus, it seems that respects only do some, but not all, of the work in explaining similarity judgements; in addition, we require an account of how information about different respects is combined to give a single similarity judgement.

While the fact that respects and hence similarity can vary does not render similarity meaningless, Goodman's argument that similarity depends on respects does have important implications for psychology. Most importantly, there will be many different similarity values between objects depending on which respects are considered. Therefore, different types of similarity can be distinguished depending on the respects in question. A number of terms have gradually, and somewhat haphazardly, been introduced to distinguish important types of similarity: *perceptual similarity* is distinguished from similarity based on conceptual properties; *global similarity* which refers to an overall comparison, underlying, for instance, the unspecified feeling that somehow, "John and Bill are very similar", as opposed to similarity centred around one or two specific respects (e.g. size); and, finally, a distinction has been drawn between *surface* and *deep* similarity. This distinction stems from the analogical reasoning literature (Gentner 1983). Here, surface similarity as based on superficial attributes is contrasted with deep or structural similarity based on common relations regardless of the mismatch of superficial attributes. A common example of such a structural correspondence is given by the similarity between Rutherford's model of the atom and the solar system. While planets and electrons do not match at a surface level, they nevertheless have corresponding roles expressed through the relation "orbit around (x, y)". Within each of these types of similarity, of course, there will be further variation, determined by the particular respects that are considered.

This flexibility of similarity has often been ignored when considering the role of similarity in the psychology of concepts. Indeed, it is widespread in the concepts literature to speak merely of "similarity" in a general way, making it necessary for the reader to work out which respects are actually under consideration. In much research, some form of perceptual similarity is assumed. Moreover, many of the criticisms levelled against similarity in the concepts literature are really criticisms of perceptual similarity (e.g. Medin & Wattenmaker 1987).

Finally, there is an alternative reply to Goodman's criticism that also sheds more light on the slightly hazy notion of global similarity. The fact that any two entities have an infinite number of properties in common also ceases to be a problem when similarity is not viewed as an objective relation between two objects but as a relation between mental representations of these objects in a cognitive agent. As mental representations must be finite, computation of similarity between objects can be thought to take place without the need for constraining respects. The crucial issue then becomes one of mental representation, of understanding what is represented and how this is selected. Arguably this is hard, but it is not arbitrary – there is a fact to the matter of what is or is not included in an agent's mental representation.

Different respects correspond to varying representations, varying either in what information is represented or in how it is weighted (or, of course, both). The two notions "similar in a given respect" and "similar in a given representation" are, from this perspective, equivalent. We find, however, that a conceptualization in terms of representation is more natural from a cognitive perspective. The perspective on global similarity then is not one of somewhat unspecified mysterious multiple respects but a comparison between two representations. The above distinctions of types correspond to differences and changes in representation, and the mechanisms of re-representation are given a prominent position both in our general understanding of similarity and of performance differences and variations in difficulty of cognitive tasks. Whatever perspective is chosen, "respects" or "representation", the problem of understanding what material enters a given similarity comparison, how this is selected and weighted, is a crucial part of understanding similarity. We will return to these issues, and the progress psychology has made here, later on in the chapter.

Having considered Goodman's critique of similarity, and suggested that similarity may, nonetheless, be a useful and explanatory notion for the psychology of concepts, we now turn to a range of models of similarity. From psychology, we consider spatial (e.g. Nosofsky 1984) and feature-based models (e.g. Tversky 1977). These models have provided the starting point for most experimental work on similarity. From artificial intelligence, we consider models of similarity used in neural networks and CBR (Bareiss & King 1989). Finally, from computer science, we consider an abstract notion of similarity based on Kolmogorov complexity (Li & Vitanyi 1993).

Spatial models

Theory

Spatial models of similarity represent objects as points in a space, with the distances between objects reflecting how dissimilar they are. These spatial representations can be viewed in two ways: merely as a convenient way of describing, summarizing and displaying similarity data or as a psychological model of mental representation and perceived similarity (Tversky & Gati 1982). In the latter view, objects are viewed as represented in an internal psychological space.[6] Objects are positioned according to their values on the respective dimensions of this space, which are viewed as the properties of the object with psychological relevance. This hypothetical space cannot, of course, be directly investigated. There is, however, a method for constructing putative internal spaces from empirical data on how similar people take objects to be. This empirical data can be of various kinds – for example, it might consist of explicit similarity judgements between pairs of objects, or data concerning how frequently people confuse each object with each of the others. According to the spatial model of similarity, similarity data, of whatever kind, can be interpreted as "proximity data" – i.e. as giving information about the distance between the objects in the internal space. Once similarity data is interpreted in terms of proximities, the problem is to reconstruct an internal space in which the distances between objects reflect, as closely as possible, the given proximity data. The problem is analogous to attempting to derive a map of a country from a table of the distances between each pair of cities. The problem of reconstructing spaces from proximity data can be solved using a set of statistical techniques known as *multi-dimensional scaling* (MDS; Shepard 1980, 1987). By using MDS, a spatial representation can be generated in which the distances between objects correspond as closely as possible to the similarities between objects.

Formally, a traditional MDS-derived model is given by:

$$d_{ij} = [\Sigma_m \; |x_{im} - x_{jm}|^r]^{1/r} \; (2.1)$$

where x_{im} is the psychological value of exemplar i on dimension m; value of r defines the distance metric. A value of $r = 2$ defines the metric as Euclidean (i.e. the shortest line between two points).[7] Other values and thus metrics are possible; $r = 1$, for example, specifies the so-called city-block metric which has also been successfully employed (here the distance between two items equals the sum of their dimensional differences). In general, it seems that it depends both on the stimulus and subject's strategy which value best fits the data (Goldstone 1994a). Stimuli with integral dimensions, that is, dimensions which are perceived together (such as hue, saturation, and brightness for colour) seem better modelled with $r = 2$, whilst stimuli with separable dimensions (size and colour, for instance) are better captured with $r = 1$ (for discussion see Nosofsky 1988; for an overview of differences found between separable and integral stimuli, see Tversky & Gati 1982).

When the spatial approach is used as a psychological model, similarity is often taken not to correspond directly to distance, but is assumed to be an exponential decay function of distance. That is, distance, d_{ij}, is converted to similarity, η_{ij}, via:

$$\eta_{ij} = \exp(-c \cdot d_{ij}^{p}) \; (2.2)$$

where c is a "general sensitivity parameter"; value of p defines the similarity gradient ($p = 1$ exponential, $p = 2$ Gaussian; Nosofsky 1988). This function is known as the *Universal Law of Generalization* and has been shown to capture the similarity-based generalization performance of both humans and a variety of animals on a range of data sets from colours to morse code signals with striking accuracy (Shepard 1987).

The similarity space on its own does not explain data from cognitive tasks – it must be supplemented with an account of how the similarity space is used. Nosofsky (1984) has developed an account of how similarity spaces could be used in a number of contexts, based on his "Generalized Context Model", which is an extension of Medin & Schaffer's (1978) exemplar account of categorization. This account has successfully fitted subject's performance on recognition, identification, and categorization tasks (e.g. Nosofsky 1988). Relating recognition, identification and categorization results, here, also requires an additional process of selective attention, to capture the fact that subjects focus on different aspects of the stimuli on each of the tasks. This is modelled through additional, flexible, weight parameters on each of the dimensions:

$$d_{ij} = [\sum_m w_m \; |x_{im} - x_{jm}|^r]^{1/r} \; (2.3)$$

in which w_m is the "attention weight" given to dimension m $(0 \leq w_m; \sum w_m = 1)$.

An increase in w_m "stretches" the space along the mth dimension, hence increasing the effect of differences on this dimension on overall similarity; correspondingly, reducing w_m "shrinks" the space on this dimension, making mismatches on this dimension less important. For illustration of this effect one can imagine a graph plotting points according to their value on the x and y axes – e.g. doubling the units per value on the x axis (i.e. 2 inches between levels on x instead of 1) will draw points further apart in the direction of this axis, thus increasing the distance between them. Given this additional mechanism of distorting the space, recognition, categorization and identification performance on this task can be related through an underlying psychological space, which is modified through attention according to the task demands.

The constraint that the weights sum to 1 offers a simple solution to a basic flaw of the unweighted spatial model. The latter fails to incorporate the effect of adding common properties to two stimuli. Intuitively, if two stimuli are modified by adding the same property to each, their similarity should increase. Dimensions, however, on which stimuli have identical values mathematically do not affect the distance between the two, as this is based on dimensional differences only. As far as these two stimuli are concerned this dimension might as well not be represented. This means one could continue to add identical properties to both stimuli indefinitely without this affecting their overall similarity – an intuitively implausible assumption which has also been experimentally invalidated by Gati & Tversky (1982).

This problem arises because the spatial model takes no account of the total number of dimensions of the representations of the objects that are being compared. If two objects are represented by three dimensions, and differ widely on all three, it seems reasonably to assume that they should not be judged as similar. If, on the other hand, they are represented by 10,000 dimensions, and differ only on these three, then it would be reasonable that they are judged to be highly similar. Intuitively, similarity is concerned with the proportion of the properties shared relative to all the properties considered. Spatial models, in their basic form, do not take account of this. By introducing attention weights that must sum to 1, Nosofsky deals with this difficulty, because adding a dimension now implies that the dimension weights for the extant dimensions are reduced. This means that they "shrink", and, hence, the impact of mismatches along the old dimensions is reduced; the new common property, as before, has no impact. The final result is a greater similarity overall.

As a psychological model, these spatial representations of similarity are of additional interest through an emerging link with neural networks. Nets, as will be discussed in more detail below, provide a very simple architecture for storing items in such a way that related items are clustered near or less near to an exemplar, depending on their degree of similarity. The items so stored likewise define a "similarity space" in the network and distance from the prototypical exemplar defines a similarity metric (Churchland & Sejnowski 1992).

Despite the appeal of the spatial approach, in particular its success in fitting a fairly wide range of data, it has come under considerable theoretical and experimental attack.

Problems

The assumptions underlying spatial models of similarity have been criticized, in particular in the work of Tversky, both on theoretical and experimental grounds (Tversky 1977, Tversky & Gati 1978, 1982, Gati & Tversky 1982, 1984, Tversky & Hutchinson 1986). Specifically, it has been argued that the continuous dimensions used by spatial models are often inappropriate, and that spatial models make assumptions about similarity that are not experimentally justified. We consider these issues in turn.

Continuous dimensions

Tversky (1977) argues that dimensional representations used by spatial models do not seem appropriate in many cases. He argues that it is more appropriate and natural to represent, for instance, countries or personality in terms of qualitative features (i.e. something an object does or does not have) rather than in terms of quantitative dimensions. This does not present a decisive argument as MDS and spatial models do not necessarily require continuous dimensions – discrete dimensions are possible and the representation of binary "features" does not automatically present a difficulty (Nosofsky 1990). On the cognitive side, conceptual stimuli might often be structured in a way that gives rise to hierarchical featural groupings or clusters and thus "pseudo-dimensions" (Garner 1978). To take an example of Rosch (1978), an automatic transmission can be treated as a feature that an object has or does not have; once it is decided, however, that the relevant set of objects are cars and that cars must have a transmission, "automatic" and "standard" become two levels on the pseudo-dimension "transmission". This also indicates that dimension vs. feature might be a processing decision that depends on task and occasion (Rosch 1978). Continuous dimensions do, however, have in principle limitations when it comes to nominal variables with several levels: there is no apparent way in which, for instance, "eye colour" which might take on the values blue, green, brown, etc. can be represented, as the different values admit of no meaningful serial ordering, a constraint demanded by the notion of dimension.

Perhaps an even more serious difficulty with representing objects as points in space is that similarity may reflect not just the collection of attributes that an object has, but the relationships in which those attributes stand, as we noted above. Representing such relationships appears to require structured representations of objects, rather than representing objects as unstructured points in space. We shall see that this problem is not limited to spatial models, but also arises for a number of other models of similarity. This problem will be discussed in detail in the context of feature-based models below.

Invalid assumptions. At the core of spatial models is the notion that similarity can be related to distance in space. Distances, by definition, must be non-negative quantities that obey the so-called metric axioms:

1. Minimality: $d_{ab} \geq d_{aa} = 0$

2. Symmetry: $d_{ab} = d_{ba}$

3. Triangle inequality: $d_{ab} + d_{bc} \geq d_{ac}$

Translating back to similarity, this implies that

1. Minimality: the similarity between any object and itself is greater than or equal to the similarity of any two distinct objects.

2. Symmetry: the similarity between objects *a* and *b* must be the same as the similarity between and *b* and *a*.

3. Triangle inequality: the similarity of *a*, *b* and *b*, *c* constrains the similarity of *a*, *c*.

Symmetry has been the main focus of attack for critics of spatial models. Similes such as "butchers are like surgeons" vs. "surgeons are like butchers", which differ in meaning with respect to whom they compliment or criticize (example from Medin et al. 1993), appear to indicate that human similarity judgements need not be symmetrical. A number of experiments have demonstrated that this effect is not specific to similes, but occurs with similarity statements ("a is similar to b") and directional similarity judgements (Tversky 1977, Tversky & Gati 1978, Rosch 1978). However, it is possible that such results can be explained not by asymmetry of similarity itself, but by other aspects of the cognitive process being studied. Enhanced spatial models which additionally allow for flexible attention weights on dimensions (Nosofsky 1988), can deal with asymmetries if they are explained in terms of "focusing": the relevant dimensions and their weightings are selected by focusing on the properties of the subject of the comparison – accordingly the space is stretched along the salient dimensions of the subject. For instance, in the comparison "surgeons are like butchers" – "surgeons" is the subject, "butchers" the referent. As the selected dimensions need not be the dimensions most salient in the referent, reversing the direction of the comparison might change the result. A different solution to this problem has been sought through the incorporation of a general notion of bias into spatial models (Nosofsky 1991).

Attempts to show that the other two metric axioms are violated (Tversky 1977, Tversky & Gati 1982) have been even less conclusive. The minimality condition is difficult to investigate because the very idea of the degree of similarity between an object and itself is problematic. The triangle inequality is difficult to test, because the constraint that it places on similarity is extremely weak. Given that similarity and distance are not necessarily the same thing, this axiom does not translate into a specific claim about similarity judgements. Recall that in many models, similarity is assumed to be an exponentially decaying function of distance. According to this assumption, the triangle inequality in distance translates into a much more complex relationship between similarities. The exact nature of this relationship depends on the precise value of exponential decay, which is, of course, not known. But many other assumptions about the relationship between similarity and distance are also possible, the only obvious constraint being that smaller distances correspond to greater similarities. But only when the precise relationship between distance and similarity is specified can the triangle inequality be translated into a claim about similarity. Evidence against the triangle inequality in a constrained case has been claimed by Tversky & Gati (1982). Moreover, there is the further problem of deciding how similarity, which is internal, relates to external behaviour. Therefore, to date, it is not clear to what extent these apparent difficulties really weigh against spatial models.

The chicken and the egg

How does the spatial model relate to the chicken and egg problem with which we began? In the discussion of concepts, we argued that categorization depends on similarity, whether directly or indirectly. According to the spatial model, does similarity depend on categorization? It does because the dimensions in the internal space are assumed to have some meaningful interpretation. Suppose, for example, that faces are classified in some way using an internal space with dimensions such as nose length, eye colour, and so on. To be able to locate a particular face in this space, so that classification can begin, requires classifying the face according to length of nose, colour of eyes, and so on. Hence, similarity depends on categorization. Note that this state of affairs is independent of whether or not we have appropriate labels for these dimensions. Unless the dimensions are meaningful, it is difficult to see how the new object can be assigned a value on each of them. Moreover, even if some way of determining the appropriate location for a new object can be found, it is difficult to imagine how this might occur without determining what properties the object has – i.e. categorizing it. Spatial models of similarity thus require that the apparently circular relationship between concepts and similarity be clarified.

Feature-based models

Theory

Feature-based models, such as Tversky's (1977) contrast model, are designed to overcome the difficulties with the spatial model. The contrast model represents objects not as points in a space with continuous dimensions but as sets of discrete, binary features (note that features need not be limited to perceptual properties). Specifically, according to the contrast model, similarity is defined as:

$$Sim(i,j) = af\,(I \quad J) - bf\,(I \cap J) - cf\,(J - I) \;(2.4)$$

I, *J* are the feature sets of entities *i* and *j*. *a*, *b*, *c* are non-negative weight parameters; *f* is an interval scale and *f (I)* is the scale value associated with stimulus *i*. This model allows for the violation of all three metric axioms discussed above as being central to spatial accounts.

Basically, similarity is an increasing function of the number of shared features $(I \cap J)$ and a decreasing function of the unmatched features of both objects $(I - J,\ J - I)$. The weight parameters *a*, *b* and *c* depend on the demands of the task. In particular, varying the focus on either the distinguishing features of *I* or of *J*, that is by increasing *b* over *c* or vice versa allows the modelling of the asymmetry of directional similarity judgements ("how similar is *i* to *j*?"). For tasks which are non-directional, e.g. where the subject is asked "how similar are *i* and *j*?" similarity judgements should be symmetrical. In the model, this requires that the parameters *b* and *c* are equal.

The scale *f* reflects the salience or prominence of the various features, thus measuring the contribution of any particular (common or distinctive) feature to the similarity between the objects. The scale value associated with stimulus (object) *i* is therefore a measure of the overall salience of *i*, which might depend on, for instance, intensity, frequency, familiarity or informational content (Tversky 1977, Tversky & Gati 1978). Because *f*, *a*, *b* and *c* can be varied, the contrast model provides a family of measures of similarity, rather than a single measure.

Problems

The contrast model makes the natural prediction that the addition of common properties increases similarity. However, this has an unintuitive consequence of its own: that similarity has no inherent upper bound. The similarity between two items can be increased indefinitely by adding elements without an ultimate value for identity being approximated. In fact, (unless ruled out by definition), the similarity of an item to itself is entirely dependent on the number of features chosen to represent it, again a rather unconvincing property.

Tversky's use of binary features rather than continuous dimensions certainly avoids the difficulties that spatial models can have with discrete properties. But it simply trades one representational problem for another – now continuous dimensions, or even nominal variables with several levels are difficult to represent. Tversky suggests various representational devices which can be used to deal with such cases, albeit somewhat awkwardly. Nominal variables of more than two values can be expressed by making use of "dummy variables" (Tversky & Gati 1982), though this solution introduces otherwise meaningless features. Similarly, ordered attributes (e.g. "loudness" levels) can be expressed through "nesting", that is through the use of a succession of sets each of which is more inclusive than the preceding one, e.g. levels of loudness: as level 1 = (), level 2 = (), and so on, or "chaining" in the case of qualitative orderings (Tversky & Gati 1982).

The representational difficulties with feature-based models do not end here, however. We noted in the discussion of theory-based concepts above that it has been argued that concepts cannot be viewed as mere collections of features. Rather, the relationships between these features must be represented, specifying the relationship of the beak, the eyes, and the tail to the whole bird. A creature with all the right features in the wrong arrangement would not be a bird! But features, as we will see, cannot express relationships. Hence, the feature-based approach to similarity appears to be unworkable from the start. Moreover, relational properties cannot simply be ignored as irrelevant to similarity judgements. Recent experiments have demonstrated that relations play an important role in human similarity judgements (Goldstone et al. 1991, Goldstone 1994b).

The problem is equally serious for spatial models. Dimensions are no more than features with a continuous number of values – these too are unable to represent relationships. If the relationships between features or dimensions are crucial in similarity judgements, not merely the features and dimensions themselves, then both feature-based and spatial models appear to be ruled out automatically as representationally inadequate.

Our most familiar means of representing relationships is natural language. The crucial difference between natural language sentences and collections of features highlights the problem. Natural language sentences have a complex syntactic structure, which can allow a finite vocabulary to be used to express an infinitely large number of statements – language is compositional. Thus, in natural language an infinitely large set of possible relationships, between arbitrary objects, can be expressed using a finite representational system. But compositionality does not appear to be possible in featural or spatial representations (Fodor & Pylyshyn 1988, Fodor & McLaughlin 1990).

Accordingly, artificial intelligence has resorted to a variety of compositional, language-like representational systems, such as *semantic networks* (Collins & Quillian 1972), *frames* (Minsky 1977), *schemata* (Schank & Abelson 1977) and various kinds of visual "sketch" (Marr 1982) in order to store relational information. In psychological terms, such a language-like, structured representation is described as a *propositional code* (Pylyshyn 1973) or *a language of thought* (Fodor 1975).

A mere collection of features is not a language; and neither is a point in a continuous space. So if objects are mentally represented using structured, language-like representations, then neither featural nor dimensional views of similarity will be sufficiently general to be satisfactory. Both approaches require some alternative way in which relationships can be represented using features or dimensions. However, no viable proposals have been put forward, and there are in principle arguments that appear to show that this is not possible (Fodor & Pylyshyn 1988, Fodor & McLaughlin 1990).[8] The problem of representing relations appears, then, to pose a serious problem for both the psychological models we have considered.

The chicken and the egg

Feature-based views of similarity also share with spatial models their status with respect to the chicken and egg problem. They confront it head on, because features are just concepts by another name. It is no help in this context to argue that these features are different, simpler, concepts than those that were originally to be explained. This provides a solution only if it is possible to arrange concepts in a hierarchy from complex to simple, where the simplest concepts/features are directly given by the perceptual system. This existence of such a hierarchy presupposes a crude empiricism, which has long been rejected as philosophically and psychologically indefensible (Fodor & Lepore 1992). The ways in which this paradox might be resolved will be investigated in the final section of this chapter.

Similarity in neural networks

Theory

Having discussed the major psychological accounts of similarity, we now turn to two important computational ideas which can be used to model cognition, neural networks and CBR. Although these computational approaches are not directly concerned with providing an account of similarity, similarity is central to the way they operate.

Neural networks (alternatively called parallel distributed processing [PDP] models or connectionist models) are a class of computational systems inspired by aspects of the structure of the brain. They consist of large numbers of simple numerical processing units that are densely interconnected, and which operate in parallel to solve computational problems. The relationship with real neurons and synapses is a loose one (Sejnowski 1986) and, within cognitive science, neural networks are generally used as cognitive models without detailed concern for neurobiological issues (Chater & Oaksford 1990). Neural networks have been applied to a range of cognitive domains including speech perception (McClelland & Elman 1986), visual word recognition (Seidenberg & McClelland 1989), learning the past tense of English words (Plunkett & Marchman 1991) and aspects of high-level cognition, including knowledge representation and categorization. For introductions into this ever-growing field the reader is referred to one of the many introductory articles or textbooks available (McClelland et al. 1986, Bechtel & Abrahamsen 1991, Churchland & Sejnowski 1992, Rumelhart & Todd 1993). Here, rather than attempt to provide a full introduction to neural networks, we shall conduct the discussion at a general level, referring the reader to the literature for further details.

One distinctive aspect of neural networks is their ability to learn from experience. A network can be trained to solve a problem on a series of examples, and will then, if all goes well, be able to generalize to novel cases of the problem to which it has not yet been exposed. A central question in neural network research concerns how this generalization occurs.

Suppose that a neural network is trained on a categorization task (unless indicated otherwise, the network we consider is a standard feed-forward network, with one layer of hidden units, trained by some variant of backpropagation. Many of the points we make apply more generally). That is, the inputs to the network are a set of examples that are to be classified, and the output of the network is to represent the category into which the current input falls. Training involves showing the network examples where the category is specified by the modeller. The network is then tested by presenting new examples and seeing whether they are classified appropriately.

The trained neural network can be viewed as a model of categorization, which, in a sense, presents an alternative to the prototype or exemplar views. Interestingly, neural networks appear to combine some aspects of both views (Rumelhart & McClelland 1985): if a network is trained on a number of distorted examples of a prototype, and then shown the prototype itself, it will classify that prototype as a particularly good example of category (i.e. the output of the network will be particularly high for that category). This is the classic prototype effect (Posner & Keele 1970). On the other hand, neural networks also appear to be sensitive to the specific examples on which they are trained – the classic exemplar effect (see Whittlesea, this volume).

Like prototype and exemplar theories of concepts, neural network categorization depends on similarity (Rumelhart & Todd 1993). But the behaviour of neural networks need not always depend directly on the similarity of the input representations – neural networks are able to form their own internal representations, on the so-called "hidden units". Classification in neural networks is therefore best thought of as determined by similarity in the internal representations of the network – thus similarity in neural networks is flexible because the internal representations are determined by the network itself, in order to provide the best way of solving the problem it has been trained on. Furthermore, each part of the internal representation used by the network need not be treated equally – some parts of the representation may be more strongly "weighted" than others (in the context of a standard feed-forward network with a single layer of hidden units, this has a very direct interpretation in terms of the magnitude of the weights from each hidden unit to the output layer).

How do the internal representations over which similarity is defined in neural networks relate to the representations used by spatial and feature-based models of similarity? Again, neural networks provide a curious combination of aspects of two different views. The internal representations consist of a set of n hidden units, each associated with a numerical value (typically between 0 and 1), its level of activation. The representation associated with the units can therefore be thought of as a point in a continuous n-dimensional space, in which each dimension corresponds to the activity level of each hidden unit. This seems compatible with spatial models of similarity. On the other hand, however, many trained neural networks learn to use a binary (or nearly binary) representation, in that the hidden unit values associated with patterns only take extreme values (i.e. almost 0 or almost 1). In such cases, the neural network can be viewed in terms of binary features, in line with feature-based models of similarity.

These remarks should be enough to suggest that neural networks provide potentially flexible and powerful models of at least some aspects of similarity and categorization, suggesting new perspectives on many issues. Researchers have attempted to exploit the potential of neural networks in a variety of ways (Shanks 1991, Gluck 1991, Kruschke 1992, Hinton 1986, McRae et al. 1993), and it remains to be seen which of these approaches will prove to be the most fruitful.

Problems

Perhaps the most significant area of difficulty for neural network models concerns the representation of relational information. This issue is vast and highly controversial, because it is central to the general debate concerning the utility of neural networks as models of cognition (Chater & Oaksford 1990, Fodor & Pylyshyn 1988, Smolensky 1988). Devising schemes for structured representations in neural networks is a major research topic as they are necessary not only in the context of categorization but for the modelling of language and large areas of reasoning. Numerous approaches have been put forward (e.g. Smolensky 1990, Shastri & Ajjanagadde 1992, Pollack 1990), of which none is wholly satisfactory. The question, thus, remains open.

Another source of problems concerns adapting similarity judgements to take account of "theory-based" effects on similarity judgements. Any effects of background knowledge will be difficult to deal with, because neural networks typically have no background knowledge – their knowledge is restricted to the category instances on which they have been trained. This again, constitutes an important research area, but is at present still in the very early stages (see e.g. Busemeyer et al., this volume; Choi et al. 1993, Tresp et al. 1993, Roscheisen et al. 1992).

If and how neural networks manage to cope with these problems remains to be seen. They indicate limitations for current network models both of similarity and conceptual structure; any final judgement, however, must be deferred.

The chicken and the egg

Neural networks offer a range of possible perspectives on the chicken and egg relationship between concepts and similarity. One picture mirrors the above discussion for spatial and feature-based models of similarity. The patterns in the inputs and outputs of neural networks can typically be interpreted. For example, the input to a word recognition model might be in terms of perceptual features at different locations, each coded by one or more units in the input to the network. This input itself, like the dimensions in the spatial models, and the features in the contrast model, therefore presupposes a classification. Here, neural networks are nothing new.

Another possibility is that similarity and concepts are mutually constraining, and that neither presupposes the other. This possibility is illustrated (though not using concepts and similarity) by neural networks which involve interactive activation. An example are interactive activation models of word recognition, in which letters and words are recognized simultaneously, so that there are mutual constraints between them (McClelland & Rumelhart 1981). Various tentative hypotheses about which letters are present each reinforce the tentative hypotheses about which words are present with which they are consistent, and inhibit those with which they are inconsistent. At the same time, reinforcement and inhibition flow in the reverse direction from hypotheses about words to hypotheses about letters. The system is designed to settle into a state which simultaneously satisfies these constraints as well as possible. Thus, in an interactive reading system, decisions about which words and letters are present are interdependent. Paradox is avoided, because there is no attempt to recognize words before letters are recognized, or vice versa. Instead, both problems are solved together. It is not yet clear whether a similar approach could be used to provide neural network models which simultaneously calculate similarity judgements and categories, subject to mutual constraint.

There is also a further possibility: that similarity and concepts emerge from a more basic process – given by the way in which the neural network learns from exposure to individual category instances. The way in which the neural network learns will determine both how categorization is carried out and the similarities between individual items. Because the behaviour of neural networks is strongly determined by similarity over the hidden units, this means that, as the network learns, that is, as it learns to form suitable categories over the hidden units, classification and similarity will inevitably be intertwined.

While these various perspectives are suggestive, it is important to stress again that the neural network approach to similarity is still underdeveloped. It is currently difficult to assess to what extent potentially promising directions provided by neural networks will ultimately prove fruitful.

Similarity in case-based reasoning

Theory

Case-based reasoning (CBR) is a computational method in artificial intelligence, from a somewhat different research tradition than neural networks. It is closely linked to both the construction of expert systems and to research on machine learning (see for overviews DARPA 1989, Slade 1991, or Kolodner 1992).

The fundamental idea of CBR is, as the name suggests, that reasoning can be based on past stored cases, rather than on complex chains of inference from stored abstract rules. It therefore requires that past cases relevant to a new situation can be retrieved successfully, and that these cases can be used to guide thinking in the new situation appropriately. Which past cases should be consulted in dealing with a new situation? The cases that are relevant are those that are similar to the new situation. Of course, the notion of similarity may vary depending on the goals and context of the reasoner, in the ways that the discussion of Goodman above suggests. So, if we are reasoning about the just outcome of legal cases, then similarity in matters of legal significance, rather than in the date and place of birth of the people involved in the trial, will be important in determining similarity. If, on the other hand, we were attempting to predict the outcome of the cases by astrological means, the birthdates might be central and the legal details peripheral in determining similarity.

Within artificial intelligence, interest in CBR has been fuelled by the recognition that rule-based approaches to the representation of knowledge encounter severe difficulties. Rule-based systems for representing information presuppose the existence of "strong domain theories" (Porter et al. 1990), that is, theories consisting of facts and abstract rules from which all required solutions can be deduced. But such strong domain theories are rarely available (Oaksford & Chater

1991) – in real-world contexts, all rules, or sets of rules, however elaborate, succumb to countless exceptions. CBR offers an attractive way out of these difficulties. Rather than having to patch up rules with endless sub-rules, to capture endless awkward cases, reasoning takes cases as the starting point. This is appealing not only as a means of building practical artificial intelligence systems, but also a framework for understanding cognition.[9]

CBR is similar in spirit to the exemplar view of concepts – large numbers of examples/cases are stored, and used to deal with the current situation. CBR is much more general, however, in three ways. First, it is concerned with reasoning of all kinds, and not simply with categorization. Secondly, cases in many CBR systems are complex structured representations, rather than points in a space or bundles of features. Therefore, CBR tackles the problem of relational properties by defining a similarity measure over structured representations. Thirdly, many CBR systems make use of prior knowledge such as general knowledge of the domain and explanations of previous cases. Hence, these systems also embody the theory-based view (Porter et al. 1990, Branting 1991).

Approaches to similarity in CBR are too diverse (Bareiss & King 1989) to be described as constituting a theory of similarity – rather, CBR provides a range of accounts, many of which may be of interest in a psychological context. In some systems, similarity requires little or no computation but is implicitly given in the way cases are represented in memory (e.g. Bayer et al. 1992). In others, explicit similarity metrics are used. Here, we find different approaches depending, in particular, on whether cases can be represented exclusively through a set of numerical values or whether symbolic representations are required. Numerical values correspond to "dimensions", which allows Euclidean distance to be used in this context (Cost & Salzberg 1993). Symbolic representations for features and relations on the other hand make use of the traditional artificial intelligence repertoire of frames, scripts or graphs, mentioned above. CBR and machine learning research is, however, continuously evolving new ways of calculating similarities between instances (e.g. Cost & Salzberg 1993). Rather than discussing any of these approaches in detail, we limit our discussion to a few general points.

First, the existence of this wealth of practically useful solutions suggests that psychological theories of similarity have only explored a small range of possible ideas about similarity.

Secondly, paradigms such as CBR and machine learning in general can provide what might be called a problem perspective, that is an understanding of similarity which originates not from high-level considerations about supposed psychological plausibility but from the need to solve a particular problem. Where these problems concern cognitive tasks generally performed by humans such a perspective can greatly contribute to our understanding.

Finally, CBR systems, while subject to limitations of their own (see e.g. Hahn & Chater 1996), provide important empirical support in evaluating models of similarity in real-world domains. While much work remains to be done before conclusions on existing algorithms can be drawn, CBR has already contributed to our understanding of the problem both by highlighting the crucial role of representation and the role of knowledge in matching two items (Porter et al. 1990), illustrating, for example, the need for structured representations, which go beyond current psychological models.

Problems

Approaches to similarity from the point of view of CBR are too various, and also too undeveloped from the point of view of psychological modelling, for a coherent list of problems to be formulated. Individual similarity metrics deserve scrutiny comparable to those provided for spatial models and the contrast model, but this would take a paper of its own. If a general weakness can be claimed, it stems from the fact that the flexibility, the context sensitivity and dependency on goals and tasks that characterizes similarity in human cognition are difficult to achieve. CBR systems are often extremely rigid in what information about a case is represented (Hahn & Chater 1996). Once an initial decision by the modeller has been made, new types of information such as previously unanticipated, novel features cannot be included (a prominent example here is HYPO, Ashley 1990). This, incidentally, is a problem they share with neural networks. In contrast to networks, exceptions to this strait-jacket of uniform representation can be found as well (an example here is given in PROTOS, Porter et al. 1990), although the flexibility of the human cognitive system remains a goal yet to be attained.

It is important to stress, however, that the reason these weaknesses, both of neural networks and CBR systems, become an issue at all, is because they are attempting to tackle a far greater proportion of the job. Spatial models and the contrast model can, because of a considerable number of suitable free parameters, fit much (possibly all) of the flexibility exhibited in human similarity judgements. They do this exclusively, however, by providing *post hoc* fits to the data. The computational models we are discussing here, in contrast, actually attempt to solve the task in question. The vital issue in similarity of what is to be represented and how the individual factors are weighted must thus be addressed in the design and implementation of these models and cannot be left to *post hoc* analysis. These questions, and the answers experiments and modelling have so far provided, will be treated in more detail in the section on feature selection and weighting below.

The chicken and the egg

With respect to the chicken and the egg, CBR offers nothing new. All systems will have a set of basic categories – features, relations, attribute values – with which they operate. Depending on the system, this set can or cannot be extended, possibly also allowing current categories to be further decomposed. At any given point in time, however, some set of categories over which similarity is computed will be treated as given.

Kolmogorov complexity

Theory

We have considered two psychological theories of similarity, spatial and feature-based models, and also the way in which similarity arises in two computational mechanisms, neural networks and case-based reasoners. We now consider an account of similarity which has a computational origin, but which is not specific to any particular computational mechanism. This account has been developed within a branch of computer science and mathematics known as Kolmogorov complexity (see Li & Vitanyi 1993, for a comprehensive introduction). Related ideas are discussed under the headings minimum message length (Wallace & Boulton 1968), minimum description length (Rissanen 1989), and algorithmic complexity theory (Chaitin 1987).

The fundamental idea of Kolmogorov complexity theory is that the complexity of any mathematical object x can be measured by the length of the shortest computer program that is able to generate that object. This length is the Kolmogorov complexity, $K(x)$ of x. The class of objects which can be given Kolmogorov complexities is very broad, including numbers and sets, but also computer programs themselves, and, more generally, representations of all kinds. Anything that can be characterized in purely formal, mathematical terms can be assigned a Kolmogorov complexity. A physical object, such as a chair, cannot, of course be generated by any computer program – and hence Kolmogorov complexity cannot measure the complexity of physical objects. But a representation of a chair (e.g. as a set of features, a point in an internal space, or using a structured representation of some kind) can be assigned a Kolmogorov complexity. An immediate query is that surely the length of the shortest program to describe an object will depend on the nature of the programming language that is being used. This is quite true, although there is a remarkable mathematical result which states that the difference between the Kolmogorov complexities given by different programming languages can differ by at most some constant factor, for any object whatever. This means that, in some contexts at least, the specific programming language under consideration can be ignored, and Kolmogorov complexity can be treated as absolute. Kolmogorov complexity, while easy to define, turns out to have a large number of important mathematical properties and areas of applications, including inductive inference and machine learning (Solomonoff 1964, Wallace & Boulton 1968, Rissanen 1989).

Kolmogorov complexity can be generalized slightly to give a notion of the conditional Kolmogorov complexity, $K(x \mid y)$, of one object, x, given another object, y. This is the length of the shortest program which produces x as output from y as input. Suppose, for example, that x represents the category "chair," and that y represents the category "bench." $K(x \mid y)$ will be low, because it is presumably relative easy (i.e. only a short program is required) to transform one representation to the other. This is because many of the aspects of the two representations will be shared, since they have many of the same properties. In particular, the length of the program needed to generate a chair representation from a bench representation will be considerably shorter than length of program required to generate the chair representation from scratch – that is, $K(x \mid y) < K(x)$. On the other hand, if chair must be derived from, say, whale, then there will presumably be no saving at all in program length – since there are no significant shared aspects of the representation which can be carried over between chair and whale representations.

The intuition is, then, that the conditional Kolmogorov complexity between two representations (i.e. the length of the shortest program which generates the one given the other) will depend on the degree of similarity between those representations. But it is possible to turn this observation around, and use conditional Kolmogorov complexity as a measure of dissimilarity. This gives a simple account of similarity, with a number of interesting properties:

(a) There is a well-developed mathematical theory in which a number of measures of similarity based on conditional Kolmogorov complexity are developed and studied (Li & Vitanyi 1993).

(b) Perhaps most importantly, this account applies to representations of all kinds, whether they are spatial, feature-based or, crucially, structured representations. Indeed, it can be viewed as a generalization of the featural and spatial models of similarity, to the extent that similar sets of features (nearby points in space) correspond to short programs.

(c) The fact that similarity is defined over general representations allows great flexibility, in that goals and knowledge of the subject may affect the representations which are formed. As with the featural model, this flexibility has both advantages, in terms of accounting for the flexibility of people's similarity judgements, and disadvantages, from the point of view of deriving testable empirical predictions.

(d) Self-similarity is maximal, because no program at all is required to transform an object into itself.

(e) The triangle inequality holds. The shortest program which transforms z to y concatenated with the shortest program which transforms y to x, is always at least as long as the shortest program that transforms z to x.

(f) It builds in the asymmetry in similarity judgements: $K\ (x\,|\,y)$ is not in general equal to $K\ (y\,|\,x)$. This asymmetry is particularly apparent when the representations being transformed differ substantially in complexity. Suppose that a subject knows a reasonable amount about China, but rather little about Korea, except that it is "rather like" China in certain ways. Then transforming the representation of China into the representation of Korea will require a reasonably short program (which simply deletes large amounts of information concerning China which is not relevant to Korea), while the program transforming in the reverse direction will be complex, since the minimal information known about Korea will be almost no help in constructing the complex representation of China. Thus, we would predict that *K (China | Korea)* should be greater than *K (Korea | China)*. This is observed experimentally (Tversky 1977).

(g) Background knowledge can be taken into account by assuming that this forms an additional input to the program that must transform one object into another. Background knowledge may radically affect the program length required to transform two objects. Whether the effects of background knowledge on human similarity judgements can be modelled in terms of the effects of background knowledge on this program length is an interesting subject for future research.

Measures of similarity based on conditional Kolmogorov complexity have yet to be developed as potential psychological accounts of similarity. This promising direction may be an important avenue of future research.

Problems

Conditional Kolmogorov complexity appears to have a number of difficulties. First, it is psychologically unrealistic, because the general problem of calculating the conditional Kolmogorov complexity between two objects is provably uncomputable (Li & Vitanyi 1993). Psychological judgements of similarity could, however, be based on crude estimates of conditional Kolmogorov complexity, of which a number are available (Rissanen 1989).

Secondly, as in simple, unweighted, spatial models of similarity, conditional Kolmogorov complexity makes somewhat bizarre predictions as common features are added to the representations of the objects being compared. Indeed, similarity decreases as more similar features are added. This suggest that some modification of the approach is required to model human similarity judgements. One obvious suggestion is that the relevant measure of dissimilarity should be given by $K\ (x\,|\,y)\ /\ K\ (x)$. This gives the prediction that objects are judged to be more similar as more and more similar features are added, but it has implications for the other properties discussed above. Whether a measure which is appropriate overall can be found is a topic for future research.

Thirdly, the approach may be insufficiently flexible. Given two representations of objects, it simply gives a global similarity value between those representations. There is no scope for weighting some aspects of the representations more highly than others, or focusing only on sub-parts of the representations. This difficulty can, perhaps, be overcome if it is assumed that the flexibility of similarity in response to changing knowledge and goals is a reflection of the flexibility of the representation of objects in the light of these factors.

The chicken and the egg

This approach appears to break out of the vicious chicken and egg circularity in a radical way. Similarity is measured in terms of program length, which makes no reference to concepts – and hence there is no circularity to explain away. But this is misleading: similarity is defined over representations of the objects being compared; and how an object is represented depends on how it is categorized. It is therefore not clear that Kolmogorov complexity provides any new insights in the apparent interdependence of concepts and similarity.

Feature selection and feature weighting: choosing respects

Leaving our introduction of models and their particular problems behind, we return to the discussion at the beginning of this section, resuming the issue of respects. There, we noted that similarity is relative to respects, rather than an absolute notion. This, we saw, is equivalent to stating that similarity is representation dependent. Fixing the respects for a given similarity comparison can, hence, be described as selecting and, possibly, weighting the factors of interest. How is this reflected in the different accounts of similarity we have described? (For a more thorough account than can be provided here, see Hahn & Chater 1996.)

The short answer is that the feature selection and weighting process is very much outside the scope of all models discussed. Neural networks and CBR systems can capture some of this process. For all other models, it is simply not addressed. The contrast model does not describe how features are chosen, but simply assumes that they are selected from a rich mental representation of the objects concerned, in the light of the task at hand (Tversky 1977). Similarly, spatial models use MDS to establish retrospectively which dimensions were of what importance to a given subject. As a tool for data analysis, this is of utility and importance. As a model of similarity it falls rather short of the mark, given that the selection of factors over which similarity is assessed is the most crucial determinant of similarity.

CBR systems and neural networks depend largely on the input representations chosen by the modeller. To the extent that these systems learn, however, they establish some weighting and selection of features. As we saw above, neural networks can learn their own internal representations, and so can choose the respects in which similarity is appropriately measured for the task on which they have been trained.

At present, however, no computational system exhibits the flexibility of humans. Our similarity judgements are, for instance, highly dependent on contexts, goals or purposes, as is evident not merely from the general considerations of the importance of respects, but also from empirical studies (Roth & Shoben 1983, Sadler & Shoben 1993, Lamberts 1994, Barsalou 1982, 1983, 1987).

Why do current computational systems not mirror this flexibility of human judgements? The answer is, because it is so hard. At a general level, respects appear to be chosen according to whether they are relevant. The general problem of determining relevance is one of the most difficult questions in cognitive science and artificial intelligence (Oaksford & Chater 1991, Chater & Oaksford 1993). Accounting for the features selection and weighting process is, thus, a tall order.

However, experimental investigation has identified a number of factors affecting both selection and weighting, which seem to arise from the way the cognitive system processes similarity judgements (Medin et al. 1993). For example, adding common features, as opposed to relations, to a pair of objects, leads to a greater increase of similarity if common features (as opposed to relations) already dominate in this pair, and vice versa for adding relations. The weight of common features, thus, seems to depend in part on whether two objects share more features or relations (Goldstone et al. 1991). Similarly, the time available for the judgement seems to affect systematically the weight attributed to the dimensions on which comparison is based (Lamberts 1995). Given short deadlines, subjects rely heavily on perceptual properties. With more time, formal category structure exerts the greater role. Including effects of this kind in models of similarity is a far more achievable goal than solving the general problem of what counts as relevant. At present, research on these questions has just begun (Lamberts 1995, Goldstone 1994b).

In summary, the question of which respects are relevant, and how strongly each should be weighted, is fundamental to any complete account of similarity. To the extent that this depends on relevance, the problem is very hard indeed. A more manageable task is presented by constraints arising from the similarity comparison process itself. The question of respects must, however, be a major topic of future research in the literature both on similarity and on categorization.

Conclusions: adequacy of current models of similarity

We have reviewed a range of current models of similarity, from psychology, artificial intelligence and computer science. The two psychological models, spatial and feature-based models, both have important limitations – perhaps most crucially in that they appear to be unable to incorporate relational information. Neural networks, CBR and conditional Kolmogorov complexity provide, on the other hand, an intriguing range of possible models. But these models are not fully worked out and moreover their psychological utility is unproved. Furthermore, we have seen that models of similarity typically leave out a crucial aspect of the psychology of similarity – concerned with choosing which respects, with what weighting, should enter the similarity comparison. Important goals for future research therefore include attempting to apply sophisticated computational ideas concerning similarity to provide better psychological models of similarity, and addressing the question that theories of similarity typically ignore: how respects are chosen.

In the previous section, we considered accounts of concepts. In this section, we have considered accounts of similarity. In the next, we focus directly on how similarity and categorization are related.

CONCEPTS AND SIMILARITY

In introducing the range of theories of concepts, we discussed the role that similarity plays in each. We have considered a range of accounts of similarity. We now reconsider the relationship between concepts and similarity from both a theoretical and an empirical point of view. We ask whether, or to what extent, theories of similarity are able to play the role required of them by theories of concepts. This involves three separate issues. First, we need to investigate how particular models of similarity can be integrated with particular views of conceptual structure, and where this leads to difficulties. Secondly, we must consider the experimental evidence concerning the relationship between similarity and categorization. Finally, it must be shown how the in principle difficulties presented by "the chicken and the egg" relationship might be resolved. We address each of these issues in turn.

Theoretical integration

Table 2.1 shows schematically the extent to which the various theories and models can be integrated. We will take each view of conceptual structure in turn and examine whether the models of similarity discussed above can be fitted in.

TABLE 2.1

	Exemplar	*Prototype*	*Theory*	*Rule*
Spatial	+	+	outside	+
Feature-based	+	+	outside	+
K-distance	+	+	+	-?
Networks	+?	+?	-	+?
CBR	+	+	+	+

As we see, the prototype and exemplar views can basically be reconciled with any view of similarity discussed. While similarity has a central role in either, they do not place any constraints on how similarity is assessed. The only query, for both these views, concerns their compatibility with similarity as found in neural networks. This is, as we recall, because the most standardly used network architectures blur the distinction between both views, showing both prototype and exemplar effects. In this sense, they can be viewed as extensions of prototype or exemplar accounts.

The theory-based view is somewhat less universally compatible. Only in two of the accounts, conditional Kolmogorov complexity and CBR, is some form of theory included directly in the process of similarity assessment. Theories – as some form of general, explicit, but partial knowledge – can affect similarity judgement in both spatial models and the contrast model only through the feature selection and weighting process. But these processes are, as we have seen, beyond the scope of either account. Hence, theory-based views of concepts are compatible, but cannot be integrated with current versions of the spatial and contrast models. Neural networks, at present, fare even worse with theory-based views of concepts, as there is currently no universal mechanism by which networks could represent and use background knowledge (but see Busemeyer et al., this volume). Finally, we noted above that conditional Kolmogorov complexity can be affected by the knowledge, because that knowledge can be used to identify simple transformations between objects, which would not otherwise be available. Conditional Kolmogorov complexity can therefore be used within a theory-based approach. A cautionary note is, nevertheless, required. Allowing knowledge to influence similarity does not guarantee that knowledge influences similarity in a psychologically relevant way – the question of whether conditional Kolmogorov complexity can appropriately capture the effects of "theory" in this respect, needs further investigation.

Finally, the degree to which definitions can be expressed in each of these frameworks again differs. In the spatial model a set of necessary and sufficient features (i.e. a definition) corresponds to a set of dimensions and values to which a point must have zero distance in order to be classified as a member of the category. In the contrast model, a definition becomes a set of specified features which must be shared by the object to be classified. In other words, the terms comprising the distinctive features [i.e. $bf(I - J)$ and $cf(J - I)$] vanish from the equation as irrelevant; the outcome of the comparison must correspond to the weighted total of the definition's features. Likewise, a CBR system can be made to match a set of necessary and sufficient conditions by introducing the constraint that these features be perfectly matched, it is not so clear, however, how definitions can be assimilated in the neural network approach – indeed, the more general question of whether neural networks can follow rules at all is highly controversial in the context of neural network models of language (Christiansen & Chater, 1997; Coltheart et al. 1993, Hadley 1994, Pinker & Prince 1988, Plunkett & Marchman 1991, Rumelhart & McClelland 1986, Seidenberg & McClelland 1989). Finally, the definitional view of concepts does not appear to have any place for the idea that similarity should be measured by conditional Kolmogorov complexity. Although possible connections can be imagined, such as that definitions are short descriptions of sets of objects, and that perhaps there is low conditional Kolmogorov complexity between pairs of objects which are members of such sets, it is not clear whether any account along these, or other, lines could be viable. In our discussion of the definitional view, we also mentioned that recent experimental investigations suggest an "interference" of similarity even where subjects used definitions or similar rules. These effects can be captured by all accounts of similarity as they merely require an ongoing similarity comparison to previous exemplars operating alongside a rule-based classification if we imagine that the former overrides the latter only above certain degrees of similarity match.

In summary, there are partial constraints between current theories of categorization and similarity. These constraints will become more important in modifying theories of concepts and similarity, to the extent that unified accounts of similarity and categorization are developed. We now move on to consider the experimental evidence concerning the relationship between similarity and categorization.

Interpreting the empirical evidence

Empirical evidence concerning the relationship between concepts and similarity comes from a variety of sources. The interpretation of much of this evidence is determined by the theoretical stance taken on concepts and similarity. We have, in passing, already mentioned a great deal of empirical data which can be viewed as support for an intimate connection between both: namely, the empirical evidence that appears to favour either prototype or exemplar views of concepts. Because the prototype and exemplar views assume such a direct and central relationship between concepts and similarity, evidence for these views is automatically evidence that concepts and similarity are closely associated. But we have also already considered evidence from the perspective of rule-based approaches to similarity – that similarity to past examples "intrudes" even on apparently rule-based classification (Nosofsky et al. 1989, Allen & Brooks 1991, Ross 1989, Whittlesea, this volume). Further credibility is lent to the idea that cognition might use similarity to stored examples in categorisation, and in reasoning more generally, by the comparative success of CBR within artificial intelligence. The reason for this was that, in domains without a strong domain theory, rules (or at least rules alone) simply will not work, as there is nothing to tell us what a sufficient set of rules ought to be. The vast majority of real world problems, however, arises in precisely such domains. Here, it is difficult to see what else, if not similarity to cases, could be the driving force.

The theory-based view of concepts, in contrast, has generated a range of experimental studies that appear to cast doubt on the relationships between similarity and categorization (Carey 1985, Keil 1989, Rips 1989, Wattenmaker et al. 1988, Wisniewski & Medin 1994). However, as we mentioned in reviewing these studies in our discussion of the theory-based view, these studies are not evidence that categorization is not based on similarity; they are evidence that categorization is not based on a simple and rigid notion of similarity, typically conceived as some kind of perceptual similarity. Once it is recognised that similarity need not be rigid, but may itself be influenced by the knowledge that the theory-based view emphasizes, then the necessity for tension between these experimental results and the role of similarity in categorization disappears. Nonetheless, at least one experiment has found a strong dissociation between similarity judgements and categorization – a result which does seem to be inconsistent with a direct link between the two, and hence, between similarity and concepts.

Rips (1989) provides two lines of experiments that undermine a straightforward relationship between similarity and categorization. In one line of experiments, he demonstrates that information such as variability of category members or frequency information differentially affects categorization and similarity judgements. Asked, for instance, whether a three-inch round object is more like a pizza than a quarter (the US coin) subjects prefer the quarter, while nevertheless preferring the classification as a pizza. These results can be explained, with some support from subjects' protocols, by the fact that pizzas allow far greater variability in size than do quarters, a fact which subjects seem to find selectively relevant to classification only. In a second line of experiments, subjects are presented with stories in which the superficial qualities of an animal undergo systematic transformation, creating greater surface similarity with another species. Nevertheless, classification as the original species is preferred. Though effects of the transformation on both categorization and similarity are observed, i.e. no strong dissociation takes place, the impact on similarity judgements nevertheless far outweighs that on categorization. In line with theory-based approaches, Rips argues that our knowledge of "essences" and underlying, non-surface features determines categorization, not superficial resemblance.

A further study, however, has produced results which indicate that strong dissociations between similarity and categorization occur only under special circumstances (Smith & Sloman 1994). Rips' results seem replicable only with sparse descriptions of objects, that is descriptions that contain only what they call "necessary" features with respect to some classification. For objects with descriptions combining necessary and merely characteristic features, categorization tracked similarity. Furthermore, even with sparse descriptions, Smith & Sloman found a dissociation only if subjects were also asked to explain their decisions.

These results are very much in line with the theory-based view. Similarity does play a role, where stimulus materials are sufficiently rich to allow similarity comparisons along dimensions perceived as relevant. Similarity, as stated several times before, is in no way limited to perceptual similarity as Rips suggests.

More generally, this line of experiments also points to the fact that the role of similarity in categorization may differ for different kinds of concepts. Goldstone (1994a) proposes the following ordering in terms of "grounding" by similarity: natural kinds ("dog"), man-made artefacts ("hammer", "chair"), *ad hoc* categories ("things to take out of a burning house"), and abstract schemas or metaphors (e.g. "events in which an act is repaid with cruelty" or "metaphorical prisons"). For the latter, Goldstone suggests, explanations by similarity are almost vacuous:

> an unrewarding job and a relationship that cannot be ended may both be metaphorical prisons, but this categorization is not established by overall similarity. The situations may both conjure up a feeling of being trapped, but this feature is highly specific and almost as abstract as the category to be explained (Goldstone 1994a:149).

At the other end of the scale high within-category-similarity has been shown to characterize at least basic-level objects[10] of many artefacts and natural kinds. At this level, category members share more features as listed by subjects than do the subordinate or superordinate category's members (Rosch 1975).

A slightly different strategy of dissociating categorization and similarity is presented in Rips & Collins (1993). Here, the experimenters aim to establish dissociations between typicality ratings, similarity ratings, and a judgement of the likelihood that a particular instance was a category member. This study, however, fails to provide persuasive data for a number of reasons. Most importantly, similarity is elicited by asking subjects how similar a particular instance is to its category. This is not a well-formed question; "robin" is not similar to "bird", a robin is a bird.[11] Presumably, subjects succeed in making some sense of this question, for example by reformulating it as "how similar is a robin to an average bird", or "to a typical bird", or "to other birds". In lieu of any information on what exactly it is that subjects do, there is no way that the data can be taken to be representative of similarity judgements as assumed to occur in the context of categorization. Additional worries rest on the fact that estimating the likelihood of an item being a category member is not the same as categorizing it (though probabilities may be part of this decision, see e.g. the "Generalized Context Model", Nosofsky 1988); given only the information that Linda is female, it is perfectly possible to judge how likely it is that Linda heads a multinational company. It does not, however, seem possible to categorize Linda, that is, say whether she is or is not head of a multinational company.[12] For this latter question, we simply lack enough information. Given, then, that the measures for the two central notions seem questionable, not much can be made of the results. However, the general strategy of the experiments, searching for differential effects of frequency information on similarity and categorization, does seem a suggestive avenue to pursue.

Clearly, the area requires further research. In particular, the interaction of similarity with "theories", that is prior knowledge, needs further specification. This, we think, requires not only more experimental but also computational work: it is only through the process of building explicit, rigorous models of theory-dependent categorization tasks that the exact need for, and thus role of, similarity assessment can be determined.

The chicken and the egg

We began this chapter by noting that concepts and similarity appear to stand in a "chicken and egg" relationship. Similarity appears to underlie categorization; but belonging to many of the same categories seems to be what makes objects similar. We then argued that this apparently circular relationship actually applies to theories in the psychological literature. We saw that current theories of concepts are all committed to the claim that concepts presuppose similarity, whether directly (for prototype and exemplar views) or indirectly (for rule-based and theory-based views). We then turned to the theories of similarity, and found that these are committed to the claim that similarity presupposes categorization. Spatial, feature-based, CBR and conditional Kolmogorov complexity approaches to similarity all presuppose categorization. We noted that neural network models also frequently presuppose categorization, although we suggested that this may not always be the case. So, our review of current theories of concepts leaves us with the conclusions that, according to current theories, concepts and similarity do stand in a chicken and egg relationship – each seems to presuppose the other.

If we accept that there is a circular relationship between concepts and similarity, how can paradox be avoided? We consider four possibilities.

1. Revise or abandon concepts and similarity. One approach is to attempt to revise the notions of concepts and similarity so that the circular relationship between them is removed. If this is not possible, then perhaps the notions must be abandoned wholesale. While this option cannot be ruled out, it is definitely to be used only as a last resort, given its severe implications for cognitive psychology. Once concepts are abandoned, for example, accounts of how knowledge is represented in memory, how language is produced and understood, what is the output of the perceptual system, and many more fundamental issues in cognitive psychology must be dramatically rethought.

2. Recursion. This approach is based on a solution to an even more basic problem of circularity: how a notion can be explained in terms of itself. In computer programming, the notion of recursion is often used to define concepts in terms of themselves in a harmless way. For example, the factorial function can be defined using the following relationships:

factorial (n) = (n) (factorial (n–1)) ·
factorial (0) = 1

The upper clause involves recursion – factorial is, in a sense, defined in terms of itself. Circularity is avoided because the problem of finding, for example, the factorial of 10 is reduced to the problem of finding the factorial of 9 by applying the recursive clause. But applying the clause again, it is reduced to the problem of finding the factorial of 8, and so on, down to the factorial of 0. Now that a complex problem has been reduced to a simple one, the simple problem can be solved directly, by breaking out of the recursion and applying the lower clause. The important point is that notions can be defined in terms of themselves, by successively reducing complex problems to simple ones of the same form.

Recursion applies equally well to cases in which there are two interdependent notions to be explained. As before, the important point is that the problems can successively be reduced to simpler problems of the same form. This is the solution to the original "chicken and egg" problem. Each chicken presupposes an egg; and each egg presupposes a chicken. But as evolutionary history is traced back, the ancestor chickens and eggs become simpler and simpler, until there are neither chickens or eggs to be explained at all.

There are various ways in which this approach might be applied to concepts and similarity. One of these has been discussed already in the context of models of similarity. We noted that concepts could be arranged in a list from most to least complex, and it could be assumed that similarity judgements on which a particular concept depends could only involve simpler concepts. Because recursion has to stop somewhere, some

concepts (or some similarity judgements) would have to be primitives, which are not explained further. Many other possible applications of the idea of recursion can be imagined. It is not clear whether any of these can provide a satisfactory account of concepts and similarity.

3. Mutual constraint. An alternative approach is that concepts and similarity must be calculated simultaneously by the cognitive system, so that each constrains the other. This was illustrated above, in the discussion of interactive activation neural network models.

Could this approach apply to concepts and similarity? The idea would be that concepts and similarity would be computed simultaneously in a mutually constrained way. That is, decisions about categorization would constrain decisions about similarity, and vice versa, but these constraints would operate simultaneously. This is an attractive idea, although it has not yet been explored.

4. The third factor. This approach assumes that the relationship between concepts and similarity is to be explained in terms of a third factor, which is more basic than either of them. Consider, for example, the degree to which metals conduct electricity, and the degree to which they conduct heat. These properties co-vary – better heat conductors are better conductors of electricity. This means that it is possible to judge how well a metal will conduct electricity by finding out how well it conducts heat and vice versa. But this does not lead to paradox, because neither notion should be explained in terms of the other. The right explanation is that there is a third factor, the atomic structure of the metal, which determines its conductivity for both heat and electricity. Moreover, this third factor makes it possible to explain why these two properties correlate as they do.

What might an appropriate third factor be in the context of concepts and similarity? A natural approach would be to specify some general goal of the cognitive system, perhaps maximizing expected utility or maximizing the amount of information gained about the environment. The general goal might require the cognitive system to construct categories; and moreover to determine similarity relationships between different objects. The critical challenge for any such approach is to show that the general goal requires concepts and similarity must co-vary, just as a challenge of atomic theory is to explain why conductance of heat and electricity co–vary. In the context of neural networks, the third cause could be the way in which the network learns when it encounters new instances. This learning might produce both classification and similarity as by-products of the change in the hidden unit representations, as we mentioned above.

Summary

We have considered four possible options for dealing with the circular relationship between concepts and similarity. It is not clear which, if any, of them can provide a satisfactory theory of concepts and similarity. But future research must take up the challenge of developing one of these accounts, or devising a different approach to explaining the circular relationship between concepts and similarity. If this is not done, theories of concepts and similarity remain in the perilous position of using explanations which presuppose the very notions that they attempt to explain. Understanding the interrelationship between concepts and similarity is therefore one of the most important, and urgent, problems facing research in both areas.

There is also a more mundane moral to be drawn from the close relationship between concepts and similarity: that it seems likely that the problems that make progress difficult in both areas may be the same. This suggests that it may be fruitful to study concepts and similarity at once, rather than as two separate domains.

CONCLUSION

The major theories of conceptual structure rely more or less heavily on similarity. This seems sound, given the fact that there is significant experimental evidence to support this view. Additionally, computational modelling within artificial intelligence has provided compelling support by highlighting the weaknesses of approaches which make no use of similarity. However, we have also seen that similarity is too complex and difficult a notion for it to be used as an explanatory primitive. Without a model of similarity, much of the problem has simply been swept under the carpet. This is all the more so, as no current model seems fully satisfactory. Furthermore, the difficulties are worsened by the intimate connection of similarity and concepts, which suggest that there are limits to the extent to which they can usefully be studied on their own. Nevertheless, we think, the feeling should not be one of dejection. The material we have reviewed does indicate that many constraints, both on theories of conceptual structure and on models of similarity, have emerged. In short, while no satisfactory solution has yet been found, it has become far clearer what we are looking for. We hope that the review of material in this chapter may provide some useful sources from which further research can begin, and indicate directions which it may prove useful to explore.

REFERENCES

AAAI 1993. *Proceedings of the AAAI–93 Workshop on Case-Based Reasoning*. AAAI Press.

Allen, S. & L. Brooks 1991. Specializing the operation of an explicit rule. *Journal of Experimental Psychology: General* **120**, 3–19.

Ashley, K. 1990. *Modeling legal argument – reasoning with cases and hypotheticals*. Cambridge, Mass.: MIT press.

Bareiss, R. & J. King 1989. Similarity assessment and case-based reasoning. In *Proceedings: Case-Based Reasoning Workshop*, 67–71. San Mateo, California: Morgan Kaufmann.

Barsalou, L. 1982. Context-independent and context-dependent information in concepts. *Memory and Cognition* **10**, 82–93.

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

Barsalou, L. 1987. The instability of graded structure: implications for the nature of concepts In *Concepts and conceptual development: ecological and intellectual factors in categorization*, U. Neisser (ed.), 101–140. Cambridge: Cambridge University Press.

Bayer, M., B. Harbig S. Wess 1992. *Ahnlichkeit und Ahnlichkeitsmasse In Fallbasiertes Schliessen: eine Ubersicht*. Techreport: SEKI Working Paper SWP–92–08.

Bechtel, W. & A. Abrahamsen 1991. *Connectionism and the mind*. Oxford: Blackwell.

Branting, K. 1991. *Integrating rules and precedents for classification and explanation*. PhD thesis, University of Texas, Austin.

Brooks, L. 1978. Nonanalytic concept formation and memory for instances. In *Cognition and categorization*, E. Rosch & B. Lloyd (eds), 169–211. Hillsdale, New Jersey: Erlbaum.

Bruner, J., J. Goodnow, G. Austin 1956. *A study of thinking*. New York: John Wiley.

Carey, S. 1985. *Conceptual change in childhood*. Cambridge, Mass.: Bradford Books.

Chaitin, G. 1987. *Algorithmic information theory*. Cambridge: Cambridge University Press.

Chater, N. & M. Oaksford 1990. Autonomy, implementation, and cognitive architecture: a reply to Fodor and Pylyshyn. *Cognition* **34**, 93–107.

Chater, N. & M. Oaksford 1993. Logicism, mental models and everyday reasoning: reply to Garnham. *Mind and Language* **8**, 72–89.

Choi, S., M. McDaniel J. Busemeyer 1993. Incorporating prior biases in network models of conceptual rule learning. *Memory and Cognition* **21**, 413–23.

Christiansen, M. & N. Chater 1997. Connectionism and natural language processing. In *Language processing*, S. Garrod & M. Pickering (eds). London: UCL Press.

Churchland, P. S. & T. J. Sejnowski 1992. *The computational brain*. Cambridge, Mass.: MIT Press.

Collins, A. & M. Quillian 1972. Experiments on semantic memory and language comprehension. In *Cognition in learning and memory*, L. Gregg (ed.). New York: John Wiley.

Coltheart, M., B. Curtis, P. Atkins M. Haller 1993. Models of reading aloud: dual-route and parallel-distributed processing approaches. *Psychological Review* **100**, 589–608.

Cost, S. & S. Salzberg 1993. A weighted nearest neighbour algorithm for learning with symbolic features. *Machine Learning* **10**, 57–78.

DARPA 1989. Machine learning program plan: case-based reasoning. In *Proceedings: Case-based Reasoning Workshop*, 1–14. San Mateo, California: Morgan Kaufmann.

Fodor, J. 1975. *The language of thought*. New York: Cromwell.

Fodor, J., M. Garrett, E. Walker C. Parkes 1980. Against definitions. *Cognition* **8**, 1–105.

Fodor, J. & E. Lepore 1992. Paul Churchland: State space semantics. In *Holism: a shoppers guide*. Oxford: Blackwell.

Fodor, J. & B. McLaughlin 1990. Connectionism and the problem of systematicity: why Smolensky's solution won't work. *Cognition* **35**, 13–204.

Fodor, J. & Pylyshyn, Z. 1988. Connectionism and cognitive architecture: a critical analysis. *Cognition* **28**, 3–71.

Garner, W. 1978. Aspects of a stimulus: features, dimensions, and configurations. In *Cognition and categorization*, E. Rosch, B. & Lloyd (eds), Hillsdale, New Jersey: Lawrence Erlbaum.

Gati, I. & A. Tversky 1982. Representations of qualitative and quantitative dimensions. *Journal of Experimental Psychology: Human Perception and Performance* **8**, 325–40.

Gati, I. & A. Tversky 1984. Weighting common and distinctive features in perceptual and conceptual judgements. *Cognitive Psychology* **16**, 341–70.

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

Gluck, M. 1991. Stimulus generalization and representation in adaptive network models of category learning. *Psychological Science* **2**, 50–55.

Goldstone, R. 1994a. The role of similarity in categorization: providing a groundwork. *Cognition* **52**, 125–57.

Goldstone, R. 1994b. Similarity, interactive activation, and mapping. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **20**, 3–28.

Goldstone, R., D. Medin, D. Gentner 1991. Relational similarity and the nonindependence of features in similarity judgements. *Cognitive Psychology* **23**, 222–62.

Goodman, N. 1972. *Problems and projects*. Indianapolis: Bobbs Merill.

Hadley, R. 1994. Systematicity in connectionist language learning. *Mind and Language* **9**, 247–72.

Hahn, U. & N. Chater 1996. Understanding similarity: a joint project for psychology, case-based reasoning, and law. *Artificial Intelligence Review*, in press.

Hinton, G. 1986. Learning distributed representations of concepts. In *Proceedings of the Eighth Annual Meeting of the Cognitive Science Society*. Hillsdale, New Jersey: Erlbaum.

Hintzman, D. & G. Ludlam 1980. Differential forgetting of prototypes and old instances: simulation by an exemplar-based classification model. *Memory and Cognition* **8**, 378–82.

Homa, D., S. Sterling, L. Trepel 1981. Limitations of exemplar-based generalization and the abstraction of categorical information. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **7**, 418–39.

Hunt, E., J. Marin, P. Stone 1966. *Experiments in induction*. New York: Academic Press.

Jones, C. & E. Heit 1993. An evaluation of the total similarity principle. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **19**, 799–812.

Keil, F. 1989. *Concepts, kinds and development*. Cambridge, Mass.: Bradford Books.

Kolodner, J. 1992. An introduction to case-based reasoning. *Artificial Intelligence Review* **6**, 3–34.

Komatsu, L. 1992. Recent views of conceptual structure. *Psychological Bulletin* **112**, 500–526.

Kruschke, J. 1992. alcove: An exemplar-based connectionist model of category learning. *Psychological Review* **99**, 22–44.

Lakoff, G. 1987. Cognitive models and prototype theory. In *Concepts and conceptual development: ecological and intellectual factors in categorization*, U. Neisser (ed.), 63–100. Cambridge: Cambridge University Press.

Lamberts, K. 1994. Flexible tuning of similarity in exemplar-based categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **20**, 1003–1021.

Lamberts, K. 1995. Categorization under time pressure. *Journal of Experimental Psychology: General* **124**, 161–80.

Levine, M. 1975. *A cognitive theory of learning: research on hypothesis testing*. Hillsdale, New Jersey: Erlbaum.

Li, M. & P. Vitanyi 1993. *An introduction to Kolmogorov complexity and its applications*. New York: Springer.

Marr, D. 1982. *Vision*. San Francisco: W. H. Freeman.

McClelland, J. & J. Elman 1986. Interactive processes in speech perception: the trace model. In *Parallel distributed processing*, vol. 2: *psychological and biological models*, D. Rumelhart & J. McClelland (eds), 58–121. Cambridge, Mass.: MIT press.

McClelland, J. & D. Rumelhart 1981. An interactive activation model of context effects in letter perception, part 1: an account of basic findings. *Psychological Review* **88**, 375–407.

McClelland, J., D. Rumelhart, G. Hinton 1986. *Parallel distributed processing – explorations in the microstructure of cognition*. Cambridge, Mass.: MIT press.

McRae, K., V. de Sa, M. Seidenberg, 1993. Modeling property intercorrelations in conceptual memory. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, 729–34. Boulder, Colorado: Cognitive Science Society.

Medin, D. 1989. Concepts and conceptual structure. *American Psychologist* **44**, 1469–1481.

Medin, D. L. & W. Wattenmaker 1987. Category cohesiveness, theories, and cognitive archaeology. In *Concepts and conceptual development: Ecological and intellectual factors in categorization,* U. Neisser (ed.), 25–62. Cambridge: Cambridge University Press.

Medin, D., G. Dewey, T. Murphy 1983. Relationships between item and category learning: Evidence that abstraction is not automatic. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **9**, 607–625.

Medin, D., R. Goldstone, D. Gentner 1993. Respects for Similarity. *Psychological Review* **100**, 254–78.

Medin, D. & M. Schaffer 1978. Context theory of classification learning. *Psychological Review* **85**, 207–238.

Miller, G. & P. Johnson-Laird 1976. *Language and perception*. Cambridge, Mass.: Harvard University Press.

Minsky, M. 1977. Frame theory. In *Thinking: readings in cognitive science*, P. Johnson-Laird & P. Wason (eds). Cambridge: Cambridge University Press.

Murphy, G. & D. Medin 1985. The role of theories in conceptual coherence. *Psychological Review* **92**, 289–316.

Neisser, U. 1987. From direct perception to conceptual structure. In *Concepts and conceptual development: ecological and intellectual factors in categorization*, U. Neisser (ed.), 1–20. Cambridge: Cambridge University Press.

Neisser, U. & P. Weene 1962. Hierarchies in concept attainment. *Journal of Experimental Psychology* **64**, 640–45.

Nosofsky, R. 1984. Choice, similarity and the context theory of classification. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **10**, 104–114.

Nosofsky, R. 1988. Exemplar-based accounts of relations between classification, recognition, and typicality. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **14**, 700–708.

Nosofsky, R. 1990. Relations between exemplar-similarity and likelihood models of classification. *Journal of Mathematical Psychology* **34**, 812–35.

Nosofsky, R. 1991. Stimulus bias, asymmetric similarity, and classification. *Cognitive Psychology* **23**, 94–140.

Nosofsky, R., S. Clark, H. Shin 1989. Rules and exemplars in categorization, identification, and recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **15**, 282–304.

Oaksford, M. & N. Chater 1991. Against logicist cognitive science. *Mind and Language* **6**, 1–38.

Oden, G. C. 1987. Concept, knowledge, and thought. *Annual Review of Psychology* **38**, 203–237.

Osherson, D. & E. Smith 1981. On the adequacy of prototype theory as a theory of concepts. *Cognition* **9**, 35–58.

Pinker, S. & A. Prince 1988. On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. *Cognition* **28**, 73–193.

Plunkett, K. & V. Marchman 1991. U-shaped learning and frequency effects in a multi-layered perceptron: Implications for child language acquisition. *Cognition* **38**, 43–102.

Pollack, J. 1990. Recursive distributed representations. *Artificial Intelligence* **46**, 77–105.

Porter, B., R. Bareiss, R. Holte 1990. Concept learning and heuristic classification. *Artificial Intelligence* **45**, 229–63.

Posner, M. & S. Keele 1970. Retention of abstract ideas. *Journal of Experimental Psychology* **83**, 304–308.

Pylyshyn, Z. 1973. What the mind's eye tells the mind's brain: A critique of mental imagery. *Psychological Bulletin* **80**, 1–24.

Restle, F. 1962. The selection of strategies in cue learning. *Psychological Review* **69**, 329–43.

Richter, M., S. Wess, K. Althoff, F. Maurer 1993. Presentations and posters at the First European Workshop on Case-based Reasoning. SEKI Report SR–93–12 (SFB 314).

Riesbeck, C. & R. Schank 1989. *Inside case-based reasoning*. Hillsdale, New Jersey: Erlbaum.

Rips, L. 1989. Similarity, typicality and categorization. In *Similarity and analogical reasoning*, S. Vosniadou & A. Ortony (eds), 21–59. Cambridge: Cambridge University Press.

Rips, L. & A. Collins 1993. Categories and resemblance. *Journal of Experimental Psychology: General* **122**, 468–86.

Rissanen, J. 1989. *Stochastic complexity in statistical inquiry*. New Jersey: World Scientific.

Rosch, E. 1975. Cognitive reference points. *Cognitive Psychology* **7**, 532–47.

Rosch, E. 1978. Principles of categorization. In *Cognition and categorization*, E. Rosch & B. Lloyd (eds). Hillsdale, New Jersey: Erlbaum.

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

Roscheisen, M., R. Hofman, V. Tresp 1992. *Neural controlling for rolling mills: incorporating domain theories to overcome data deficiency*. Advances in Neural Information Processing 4. San Mateo, California: Morgan Kaufman.

Ross, B. 1984. Remindings and their effects in learning a cognitive skill. *Cognitive Psychology* **16**, 371–416.

Ross, B. 1987. This is like that: The use of earlier problems and the separation of similarity effects. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **13**, 629–37.

Ross, B. 1989. Some psychological results on case-based reasoning. In *Proceedings of the Workshop on Case-based Reasoning*, San Mateo, CA: Morgan Kaufmann.

Ross, B. & P. Kennedy 1990. Generalizing from the use of earlier exemplars in problem solving. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **16**, 42–55.

Ross, B., S. Perkins, P. Tenpenny 1990. Reminding-based category learning. *Cognitive Psychology* **22**, 460–92.

Roth, E. & E. Shoben 1983. The effect of context on the structure of categories. *Cognitive Psychology* **15**, 346–78.

Rumelhart, D. & J. McClelland 1985. Distributed memory and the representation of general and specific information. *Journal of Experimental Psychology: General* **114**, 159–88.

Rumelhart, D. & J. McClelland 1986. On learning past tenses of English verbs. In *Parallel distributed processing*, vol. 2: *psychological and biological models*, D. Rumelhart & J. McClelland (eds), 216–271. Cambridge, Mass.: MIT Press.

Rumelhart, D. & P. Todd 1993. Learning and connectionist representations. *Attention and Performance* XIV, 3–30.

Sadler, D. & E. Shoben 1993. Context effects on semantic domains as seen in analogy solution. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **19**, 128–47.

Schank, R. & R. Abelson 1977. *Scripts, plans, goals, and understanding*. Hillsdale, New Jersey: Erlbaum.

Seidenberg, M. & J. McClelland 1989. A distributed, developmental model of word recognition and naming. *Psychological Review* **96**, 523–68.

Sejnowski, T. 1986. Open questions about computation in the cerebral cortex. In *Parallel distributed processing*, vol. 2: *psychological and biological models*, D. Rumelhart & J. McClelland (eds). Cambridge, Mass.: MIT Press.

Shanks, D. 1991. Categorization by a connectionist network. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **17**, 433–43.

Shastri, L. & V. Ajjanagadde 1992. From simple associations to systematic reasoning: a connectionist representation of rules, variables and dynamic bindings using temporal synchrony. *Behavioural and Brain Sciences* **12**, 456–78.

Shepard, R. 1980. Multidimensional scaling, tree-fitting, and clustering. *Science* **210**, 390–99.

Shepard, R. 1987. Toward a universal law of generalization for psychological science. *Science* **237**, 1317–23.

Slade, S. 1991. Case-based reasoning: a research paradigm. *AI Magazine* **13**, 42–55.

Smith, E. & D. Medin 1981. *Categories and concepts*. Cambridge, Mass.: Harvard University Press.

Smith, E. & S. Sloman 1994. Similarity- versus rule-based categorization. *Memory and Cognition* **22**, 377–86.

Smolensky, P. 1988. On the proper treatment of connectionism. *Behavioral and Brain Sciences* **11**, 1–74.

Smolensky, P. 1990. Tensor product variable binding and the representation of symbolic structures in connectionist systems. *Artificial Intelligence* **46**, 159–215.

Solomonoff, R. 1964. A formal theory of inductive inference, part 1. *Information and Control* **7**, 1–22.

Tanaka, J. & M. Taylor 1991. Object categories and expertise: is the basic level in the eye of the beholder? *Cognitive Psychology* **23**, 457–82.

Trabasso, T. & G. Bower 1968. *Attention in learning: theory and research*. New York: John Wiley.

Tresp, V., J. Hollatz, S. Ahmad 1993. Network structuring and training using rule-based knowledge. In *Advances in Neural Information Processing 5*, C. Giles, S. Hanson, J. Cowan (eds). San Mateo, California: Morgan Kaufman.

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

Tversky, A. & I. Gati 1978. Studies of similarity. In *Cognition and categorization*, E. Rosch & B. Lloyd (eds), 79–98. Hillsdale, New Jersey: Erlbaum.

Tversky, A. & I. Gati 1982. Similarity, separability and the triangle inequality. *Psychological Review* **89**, 123–54.

Tversky, A. & J. Hutchinson 1986. Nearest neighbour analysis of psychological spaces. *Psychological Review* **93**, 3–22.

Wallace, C. & D. Boulton 1968. An information measure for classification. *Computing Journal* **11**, 185–95.

Wattenmaker, W., G. Dewey, T. Murphy, D. Medin 1986. Linear separability and concept learning. *Cognitive Psychology* **18**, 158–94.

Wattenmaker, W., G. Nakamura, D. Medin 1988. Relationships between similarity-based and explanation-based categorization. In *Contemporary science and natural explanation*, D. Hilton (ed.). Brighton: Harvester.

Whittlesea, B. 1987. Preservation of specific experiences in the representation of general knowledge. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **13**, 3–17.

Wisniewski, E. & D. Medin 1994. On the interaction of theory and data in concept learning. *Cognitive Science* **18**, 221–81.

NOTES

1. Rather *ad hoc* seeming grouping of objects are found as so-called goal derived categories – e.g. "things to take out of a burning house" (Barsalou 1983); these will be discussed later. For "dribs", however, no such unifying goal is in sight.
2. This term is widely used to cover family resemblance (Rosch 1975, Rosch et al. 1976) and probabilistic accounts (Komatsu 1992, Smith & Medin 1981).
3. For example, prototypes can be viewed as abstractions (such as a feature list) or as a particular exemplar (Lakoff 1987). The former is exemplified in notions of central tendency as average properties, modal properties, or modal correlations of properties. Viewing the central tendency as one or a number of particularly representative instances of the category illustrates the latter approach. See Barsalou (1987) for discussion of the numerous ways in which a particular exemplar might be "typical".
4. The definitional view underlies the main psychological research on artificial concepts from 1920 to 1970 (Smith & Medin 1981). Indeed, early empirical research embodied the assumption in the choice of experimental materials used: subjects were typically asked to learn to classify artificial materials, where the "correct" classification was given by a rule formulated by the experimenter (Bruner et al. 1956, Hunt et al. 1966, Levine 1975, Neisser & Weene 1962, Restle 1962, Trabasso & Bower 1968).
5. It is possible to argue that definitions of everyday concepts might exist in an internal "language of thought", even if these definitions could not be given in natural language. While logically possible, this view is unattractive, in the absence of any concrete proposals concerning the nature of this language of thought and how definitions of everyday concepts can be framed in terms of it.
6. In a sense, the notion of psychological space is not particularly well defined: there are no commitments as to what exactly this space is, whether it is a long-term representation or not, nor whether it is explicitly similarity that is represented here or whether the representation of similarity it generates is merely a by-product of a general scheme for the representation of objects.
7. In the Euclidean case, the equation is merely a generalization of Pythagoras' theorem to any number of dimensions: the square of the length of the hypotenuse is equal to the sum of the squares of the lengths of the other two sides. Hence, the length of the hypotenuse equals the square root of this sum. The distance between two points, however, is the hypotenuse of the right-angled triangle defined by the stretch (the differences) between the values of both points on both co-ordinates as the other two sides.
8. It is, of course, true that in a sense anything can be a feature (Tversky 1977) or a dimension, and any relation can also be represented as a feature: the fact that some individual a is the mother of b can make use of the two place relation "mother-of", i.e. mother (a, b), can also be expressed with one-place predicate (i.e. a feature) "mother-of-b", that is mother-of-b (a). This, however, does not solve the problem. The choice between a 1-placed, featural and an n-placed relational representation is not arbitrary as it determines the choice of primitives in the representation of entities. This, in turn, directly affects the similarity between entities as it determines in what ways they can be compared: if representational specificity leads to

"left-eye" and "right-eye" as primitives, one cannot even compare two eyes within the same bird. The problem is one of a general tension between the need for simple features which allow comparison and the need for encoding relations between features. The situation is one of "having your cake and eating it" and it seems that it can only be avoided by using structured representations.

9. Examples of systems with a primary emphasis on cognitive modelling are to be found in Riesbeck & Schank 1989. Practical applications (e.g. fault diagnosis) can be found in the relevant conference proceedings such as Richter et al. (1993) or AAAI (1993). Examples of commercially available products are ReCall by ISoft S. A. or Remind by Cognitive Systems Inc.
10. Within a hierarchy of abstraction such as "rocking chair", "chair", "furniture", the basic level – "chair" – is that which seems cognitively privileged in the sense that it is first learned, most freely produced, first accessed, and most quickly confirmed (see, e.g. Murphy & Lassaline, this volume; Rosch 1975, Rosch et al. 1976; see also, e.g. Tanaka & Taylor 1991).
11. Rips & Collins' own reply that such questions are common in ordinary language as illustrated by questions such as "How similar is Alice to Woody Allen's other movies or how similar is Montreal to European cities?" (1993:483) misses the point as it lacks precisely the element it ought to have: "Woody Allen's other films" are instances, not a superordinate category, of Alice.
12. This holds for all three accounts of conceptual structure. We cannot tell whether, for examples, a definition of "head of multinational company" applies, for, whatever it may be, it will not contain "male" as a necessary and sufficient definition. In both the exemplar and prototype view, the lack of further detail about Linda makes the necessary similarity computation impossible; again, it need not concern us what exactly exemplars or the prototype of this category look like, because even if "male" was a specified attribute, both accounts, by definition do not require that all attributes are matched. For both, nothing follows from the existence of non-matching (non-necessary) features on their own.