CHAPTER ONE

THEORIES ABOUT CONCEPTS

1.00 CHAPTER OVERVIEW

Generally, the thesis is looking at what constitutes reliable knowledge and understanding of the world, and the objects and creatures in it. One of the issues dealt with is the source of conceptual meaning, whether it is derived from the external world or from our internal cognitive system. This first chapter looks at philosophical, semantic and psychological views on the source of concepts.

Section 1.01 of the chapter defines the nature of the problem and two general questions to be investigated by the thesis. The aims of the thesis are to show that we have the categories we do (and not others) because of the concepts we construct. The specific aim of this chapter is to assess the various answers to the question "how do concepts arise?", and their implications for the issue of stability of conceptual knowledge.

Section 1.02 gives a very brief overview of the major theories of concepts. These views include: concepts as organizational principles of reality, that is, "objective Truth"; concepts as representations of organizational principles; concepts as internal representations of the external world; and concepts as building blocks of knowledge.

A number of points are discussed. Firstly, the semantic tradition of dividing a word into its intensional and extensional components is described. An analogy is drawn between the functions of intension and concept, and between extension and category. Secondly, the metaphysical view of concepts is described. These are organizational principles of an "objective" material reality, which persons either know, or of which they are ignorant, or about which they have false beliefs (Sutcliffe, 1993). Thirdly, representational theories for conceptual knowledge are described. Concepts might be described as internal representations of the external world; whilst another approach views concepts as the building blocks of knowledge, constructed through our knowledge of society and culture.

This overview leads to section 1.03 which assesses the adequacy of these theories of concepts in accounting for conceptual stability within a person and across people. The assumptions implicit in the various accounts are discussed. If a model is put forward, where the source of meaning does not lie in the external environment, the problem becomes one of instability of meaning and the consequent need for embeddedness of concepts. Approaches from natural kinds (Schwartz, 1979) believe that objective information from the environment provides the most stability in a concept; whilst formal semantics claims that dictionary knowledge defines the true meaning of a word-concept. Similarity-based psychological theories claim that normative information is of most importance; and the epistemological approach states that both normative and experiential information is required for stability. A hypothesis is developed that proposes the best account for the stability of concepts will be provided by a model which makes use of information from both sensory experience and conceptual knowledge. Section 1.04 sums up the argument and hypotheses concerning conceptual issues and the contents of concepts.

1.01 THESIS QUESTIONS AND AIM OF CHAPTER

Rosch (1978) began her classic article, "Principles of categorization", with the following example of a taxonomy of the animal kingdom, said to be extracted from an ancient Chinese encyclopedia entitled the *Celestial Emporium of Benevolent Knowledge:*

> On those remote pages it is written that animals are divided into (a) those that belong to the Emperor, (b) embalmed ones, (c) those that are trained, (d) suckling pigs, (e) mermaids, (f) fabulous ones, (g) stray dogs, (h) those that are included in this classification, (i) those that tremble as if they were mad, (j) innumerable ones, (k) those drawn with a very fine camel's hair brush, (l) others, (m) those that have just broken a flower vase, (n) those that resemble flies from a distance (Borges, 1966, p. 108)

Rosch (1978) went on to argue that, whilst such classes might exist in the imaginative mind of a poet, no culture in the world would include them in its practical or linguistic categories of artifacts or animals. She argued that human categorization is not an arbitrary product of whimsy or accident, but is governed by the similarity structures inherent in the external world, and that these structures are perceived in roughly the same way across cultures. Lakoff (1987) did not take Rosch's attitude to this example, namely, that it was simply an imaginative classing of objects, and could not reflect how one culture might organize its world into categories which it uses everyday. Lakoff (1987) points out that Borges' passage captures the Western reader's idea of nonwestern cultures. He claims that "people around the world will categorize things in ways that both boggle the Western mind and stump Western linguists and anthropologists" (1987, p. 92).

As supporting evidence for this stance, he reports R.M.W. Dixon's (1968; 1982) research into traditional Dyirbal, a language of the North Queensland aborigines. Dyirbal precedes all its nouns by a variant of four words: *bayi, balan, balam, bala.* These words classify all objects in the Dyirbal universe, into what at first appears to be weird categories. Using this language entails that the correct classifier precede an appropriate noun, and that speakers do this in a spontaneous way, without recourse to conscious reflection. Examples of the various categories are listed below:

bayi: men, possums, bats, the moon, storms, rainbows, boomerangs, some spears, most snakes, most fishes, some birds, most insects;
balan: women, fireflies, crickets, anything connected with water or fire, sun and stars, shields, some spears, some trees; some snakes, some fishes, most birds;
balam: all edible fruit and the plants that bear them, honey, wine, cake;
bala: parts of the body, meat, bees, wind, yamsticks, noises and language, mud, stones.

The categories given above seem at least as fantastic as Borges' classes from the ancient Chinese encyclopaedia. Yet Dixon (1982) was not "stumped" by it, and made sense of it. He proposes basic schemata, which it is assumed young Aboriginal children need to learn, when acquiring competence in the language.

bayi: human males; animals;
balan: human females; water; fire; fighting; dangerous things;
balam: nonflesh food;
bala: everything not in the other classes.

Dixon (1982) has also extracted a number of principles from the linguistic lists, which would explain the inclusion of some seeming exceptions in the four classes. The principle most pertinent here, because it speaks to the argument developed in this chapter, is the "myth-and-belief" principle:

> If some noun has characteristic X (on the basis of which its class membership is expected to be decided) but is, through belief or myth, connected with characteristic Y, then generally it will belong to the class corresponding to Y and not that corresponding to X (Lakoff, 1987, p. 94).

Even the seeming exceptions to the basic schemata are understandable, once Dixon describes the conceptual belief system driving them. For example, although Westerners would classify birds as animals, they are not in the *bayi* class with other animals. Birds are believed to be the spirits of dead human females, and so are in the *balan* class. Fish are mostly in the *bayi* class with other animate beings, but the stone fish and gar fish are harmful and so are in the *balan* class. According to myth, the moon and the sun are husband and wife; so the moon is in the *bayi* class with other husbands, and the sun is in the *balan* class with other wives. Wind is in the *bala* class, but storms and the rainbow are believed to be mythical men and so are placed in the *bayi* class.

The point of providing these two examples is to illustrate three properties often found in people's category behaviour: stability, flexibility and coherence. Rosch's (1978) objection to the Chinese Emporium was that it violated the correlational structure of the world. In other words, they were not stable, everyday concepts which people would continue to use across situations and at different times, and to communicate with others. The thesis examines whether the stability of our conceptual knowledge can be based solely on the structure of physical appearances. Dixon's (1982) example is valid as long as it is assumed that Dyirbal's linguistic classes indicate categories of thought. Lakoff (1987) argues that they do, because the various classes are organized on *conceptual* principles (see Lakoff's discussion on Whorf's (1956) issue of linguistic relativity, pages 318-320). The example does illustrate the seeming lack of constraint in people's everyday categories across cultures, since the Dyirbal's linguistic categories are so different to Western ones. The two cultures do seem to share stable concepts, however, because once the "myth-and-belief" principle is explained, Westerners also can cohere the objects into the same, now comprehensible, categories. This last fact illustrates the flexibility of people's conceptual relations, which is evident in any categorization of objects. These two properties, coherence of properties and flexibility of conceptual relations, are also examined in the thesis.

The two specific terms which seem to be most used in different ways by different researchers are *concept* and *category* (for definitions of other cognitive terms, see the Glossary in Appendix A). Medin (1989) has defined them below:

> Roughly, a concept is an idea that includes all that is characteristically associated with it. A category is a partitioning or class to which some assertion or set of assertions might apply (Medin, 1989, p. 1469).

Some theorists, such as Hospers (1990), use *concept* and *category* interchangeably to mean *concept,* but most theorists consider *concept* and *category* as separate theoretical constructs (Barsalou, 1992). A person has knowledge of a concept (say, *salt*) when s/he can achieve two cognitive functions. One is the ability to distinguish *salt* from other members in the same category (say, *pepper, cumin, coriander*). Even before such distinctions can be made among the concepts, the person needs to have a concept of what constitutes *salt,* thus being able to distinguish it from *non-salt;* or at a more abstract level, distinguish *Spices* from *non-Spices.*

A general consensus exists on a concept's role in categorization: a concept is "information that allows people to discriminate members of a category from non-members" (Barsalou, 1992, p.153). Having such a concept might be more specifically defined as having a "criterion-in-mind", which enables the person to categorize an object, entity or event; and also distinguish it from all other concepts (Hospers, 1990, p.38). Thus, one should be able to categorize Fido as a *Dog* rather than a *Cat* by using the common concept of *Dogs;* and after that, distinguish Fido from other concepts in the same category, say *Great Danes, terriers* or *poodles.* The problem consists of how to explain people's conception of the individual members (*poodles, Great Danes, terriers*) as a comprehensible category, and one which can be distinguished from other categories, such as *Cats.* (Medin, 1989; Medin & Wattenmaker, 1987; Murphy & Medin, 1985).

So whilst a general consensus might exist about a concept being described as "having a criterion" which guides categorization of objects, the controversy centres on what this criterion or information might be. In other words, what constitutes reliable knowledge and understanding of the world? The contrast between Rosch's (1978) and Lakoff's (1987) attitudes to the Borges' (1966) example reflects the issue about the source of conceptual meaning; whether meaning is derived from the external world or from our internal cognitive system. This very general question subsumes two questions being investigated by this thesis (Medin, 1989). They concern:

(a) how do certain concepts arise; and
(b) why we have the categories we do, and not others.

The first question (considered in this chapter) includes philosophical, semantic and psychological views on the source of concepts, which holds implications for the kind of stable knowledge which constitutes our concepts. The second question will be considered from an empirical background, and will be addressed in chapter 2.

They are important questions because the psychological view of concepts and categories is that they serve as the components of human thought (Medin, 1989). A study of the concepts people hold about their categories will indicate the potential of the human mind. Some realist philosophers might describe concepts as abstract symbols which are the exact reflection of the categories in the environment, and thought as the manipulation of those symbols (Sutcliffe, 1993). Such philosophical views assume the human mind to be logical, its concepts a reflection of an objective "Truth" derived from the ontological categories in the environment.

Given Dixon's (1982) myth-and-belief principle, however, western logic need not be the only version of Truth. People can *impose* their own category structure on the world, not just *discover* structure which is already there. Some psychologists claim that our concepts consist of perceptions of appearance and nothing else. There is little room in such a mind for conceptual beliefs and understanding about the environment. It is also a limited view of the mind, as we can have concepts of creatures or objects which do not exist in the environment at all (e.g., the Loch Ness monster) (Medin, 1989).

So the general question of how we conceptualize our world might require an explanation which is a little more complex than one which equates categories of objects in the world with their reflections for concepts in the mind. Any answer to the two questions addressed in the thesis has implications for such issues as the mind's creativity, whether the environment is carved into objectively "true and correct" category structures, and whether appearance of the environment is the sole reality. It will be argued in this thesis that we have the categories we do, because of the concepts we construct in order to understand our experiences. Furthermore, the concepts (such as the beliefs we entertain about the world around us) have been constructed from the knowledge we hold about particular domains of experience in our individual worlds.

The aim of this chapter is to assess various philosophical, linguistic and psychological views on concepts, and to reach some conclusion about which theory of concepts might best account for how people understand the world around them in a stable fashion. The first issue, which cannot be tested empirically, concerns the source of meaning - does it lie in our heads or in external reality? The various conceptual views can be said to fall under two approaches : metaphysical or epistemological. It is suggested here that the most psychologically real view would be one which can give a complete account for the stability of conceptual knowledge. The second issue is concerned with the stability of knowledge. Do our concepts consist of innate notions of Truth, or experiential information (be it direct or indirect experience)? It will be argued in the third section that it is the kind of knowledge (objective, dictionary, normative, or subjective) contained in concepts which will contribute to their stability. The philosophical background to these issues is placed in Appendix B.

1.02 THEORIES OF CONCEPTUAL MEANING

This section looks at the various theories of concepts, most of which are based on the assumption that concepts are representational entities, rather than an objective state of affairs of which lay people might (or might not) be aware. The thesis is concerned with the psychological study of concepts and categories, and so the criticisms made of representational approaches - that they beg the question of how the world is known - will not be considered here. However, for details of these views, see Gibson, 1966; Maze, 1983, 1991; McMullen, 1988, 1991; and Michell, 1988. In the last section of this chapter, it is argued that the view of concepts as representational entities provides the best account of how people might understand their world, because it avoids certain inadequacies created by the other approaches.

1.02.1 Concepts as word definitions

The traditional theory of meaning describes a general noun as having an intension and an extension; and the role of a concept is to mediate between them (Frege, 1892). The word's intensional component specifies its formal truth conditions or meaning, by providing a list of descriptive properties which are ideally, necessary and jointly sufficient for the application of the term. In one sense of "meaning", the intension of a noun constitutes the object's meaning, a dictionary-like definition of the term. Thus, someone knows the meaning of the term *bachelor* if they have the concept of *unmarried, adult, male, human.* These properties of the concept constitute the truth conditions for the term *bachelor.* The intension of a term is often conceived of as a set of criteria for application of the term. Sometimes the intension is called the *connotation* of the term (Schwartz, 1979).

The extension of a general noun is the class of things to which the noun applies. The extension identifies a word's possible referents in the environment. For example, the extension of *Bachelors* might include such members as *Popes, homosexuals, single men, engaged men* (Barsalou, 1992; Johnson-Laird, 1983; Lyons, 1977a, 1977b). Thus, the extension is the class of things which fulfil the properties listed in the intension. Sometimes the extension of a general noun is called the *reference* of the term, and sometimes it is called the *denotation* (Schwartz, 1979).

In formal semantics, the mind-environment relationship is captured by the intension (concept) and the extension (objects-in-the environment). The mathematical philosopher Carnap saw the intension of a term as a criterion for judging whether objects in the world belonged to the term's extension, so that the intension-extension components of a word can be equated (roughly) with the distinction between *concept* and *category* (Hampton & Dubois, 1993; Putnam, 1977; Quine, 1977). The role of concepts is similar to that of a word's intension, using criterial information about what constitutes membership in the category and to judge whether certain inferences are allowable; whilst categories are concerned with the extension of the concept's meaning, the specific members in the world which fit those membership criteria or description.

The purpose of formal, logical concepts is to mentally symbolize the metaphysical structure of the world. It is the person's task to discover these *under*lying organizing principles (Anderson, 1991b; Sutcliffe, 1993). Formal semantics describes thought as the manipulation of abstract symbols (words and their mental representations), which in themselves are meaningless, and can be made meaningful in only one way: through modelling the structure of the categories in the material world, or where possible worlds are concerned, that of fictional characters like unicorns (Lakoff, 1987).

The accounts based on formal semantics of how our concepts mirror the logical structure of the environment are elegant and plausible. Firstly , their explanations for items' coherence as a comprehensible category would be relatively simple, being based upon a one-to-one correspondence between a word's intension in the mind and its extension of referent objects in the environment. Secondly, the theory's description of the intensional concept as a defining list/rule provides a precise and clearcut representation of meaning for a category. This means that, according to formal semantics, an intensional concept can represent more than one object, which saves a great deal of effort in cognitive processing and storage of conceptual information.

The main inadequacy of Frege's (1892) semantics is indeterminacy of reference, because the theory proposes a conceptual representation which consists of an abstract formula, with no details for the identification of individual members. A *Feline* could be either a *kitten* or a *tiger,* and yet again, *a Kitten* could be either *smudge* or *tabitha*. The concept represents only the underlying organizing principles of these objects in the environment, so it cannot distinguish between individual members of the same category. A fuzzy concept based upon appearance of the external environment would avoid this problem. Similarity-based concepts, especially exemplar prototypes, include characteristic features and thus allow individual members to be distinguished one from another.

1.02.2 Concepts as (lay) stereotypes

Another theory of meaning based upon assumptions of metaphysical realism is that of the philosopher Hilary Putnam. He claimed that the word-symbol (or intension) derives its meaning externally from the environment by the word's direct association with its referent object. That object is a kind of thing (or category) determined by its unseen essence. His account of direct reference and of natural kinds sees extensions as part of meaning, so that words fit the structure of the world directly (Putnam, 1975a, 1975b).

So instead of an abstract, logical concept acting as a link between intension and extension, different kinds of words (natural or nominal) have "true" meanings which originate in the external environment, not from conceptual representations. Where *natural* kind terms are concerned, the word derives its "true" meaning directly from the structure of the natural world: its association with the referent object and its natural category. Putnam originally subscribed to the doctrine of essentialism for natural kinds which states that "essential nature is not a matter of language analysis but of scientific theory construction" (1975a, p.104, but see Pylyshyn & Demopoulos, 1986, p.242). On the other hand, the meaning of *nominal* kind terms resides in their name, which receives its definition through social convention. For example, the description of a *bachelor* or *triangle* would be determined through usage and general consensus. Whereas natural kind terms are used referentially, to "pick out" their associated referent object and its natural category, nominal kind terms are used attributively, to describe the object and its category (Komatsu, 1992).

The majority of people use a stereotyped, community-based notion of a term's "true" meaning, which might well be inaccurate. This lay stereotype is derived from diagnostic definitions (by scientific experts) of the kind of thing (natural or nominal) the referent object is, and a direct description of the objects in the world to which the word correctly refers. For example, experts in precious metals have the diagnostic tests to identify the essential features in the composition of gold, and thus distinguish it from other precious metals. Their knowledge would be passed on to their social community, who might interpret it inaccurately. Lay knowledge usually takes the form of a vague, stereotyped notion of the word's "true" meaning. People use these lay stereotypes to categorize referent objects, but their mental representations can consist of vague definitions or even mistaken beliefs about the object's essence. These lay concepts do not mediate the link between natural kind terms and their referents, that is, they cannot provide the term's "true" meaning (Putnam, 1975b).

For natural kind terms, there is a division in linguistic labour (Putnam, 1977). One part of the linguistic labour is carried out by experts, who "discover" what is the object's true essential identity, for example, what makes a *cod*, a *cod* (and not a *trout).* The second part of the linguistic labour is carried out by the lay community, from which emerges a vague stereotype of a *cod* based upon these scientific definitions, and which is used to categorize a creature as belonging to the natural kind of *cods.*

Putnam's theory implies that an object can be categorized in the one automatic process of direct reference without any need for a conceptual representation to act as an intermediary between the two. Instead, to be identified, the person uses a stereotyped notion to pick-out or label an object without any cognitive understanding of the "true" or diagnostic meaning of the term. This proposed lack of any need for a person to have a conceptual representation of a "truthful" meaning is the main difference between Putnam's (1989) stereotypes and Frege's (1892) logical concepts. Their main point of agreement is that they equate meaning with objective Truth, rather than cognitive understanding (Santambrogio & Violi, 1988).

The main inadequacy of Putnam's theory (1975a; 1975b) is that it implies a rigidity of categorization. For example, people can and do categorize the same heavy object as a natural kind (for example, *a rock*), and just as easily they categorize it as a nominal kind, perhaps an artifact (*a weapon* or *a door-stop).* An object is said to have inherent meaning as a natural kind, independent of the person perceiving and identifying it, yet simple observation reveals that human categorization just does not work that way.

1.02.3 Concepts as fuzzy prototypes

Frege's (1892) semantic view of concepts and Putnam's (1975b) philosophical view both assume that the world outside of the mind is the source of "true" conceptual meaning, whose truths would continue to exist even if no human being remained on Earth. The next two approaches to meaning introduce, in differing degrees, a more human element in the mind-environment relationship.

Fuzzy concepts are so-called because they describe people's perception of the world and its categories as ill-defined and imprecise. People are said to perceive their environment in a fuzzy way, because its entities and their natural categories are linked by an overlapping of *characteristic* features; whereas the formal approach assumed that people perceived *defining* features in a precise way. Consequently, the membership structure within any fuzzy natural category is said to be graded according to the typicality of its members. Graded structure in biological concepts was first suggested by the natural philosopher William

Whewell (1847), who noted the distinctions which can be drawn between the structures of logical and biological concepts :

> The class is steadily fixed, though not precisely limited; it is given, though not circumscribed; it is determined, not by a boundary line without, but by a central point within; not by what it strictly excludes, but by what it eminently includes; by an example, not by a precept. (1847/1967, vol.1, p. 494).

As can be seen in the quotation above, with fuzzy prototypes the focus has shifted to the structure of physical appearance, whereas with logical concepts and stereotypes, the abstract rule or the unseen essence are important. In the above quotation, logical concepts focus upon *intensional criteria* such as "what it strictly excludes", "by a precept" "by a boundary line without", et cetera. With fuzzy concepts, at least the Roschean prototype ones, the focus is upon the *category extension* and the structure of its members, such as "not precisely limited", "a central point within", and "what it includes" (Brewer, 1993).

In her 1973 article, Rosch is not clear on whether the graded structure is a property of the world or a property of the human perceiver. However, it is not likely that she believed in a uniquely mental concept of gradience imposed upon a world of arbitrary instances (Brewer, 1993).Unlike the strongly metaphysical approach of logical concepts, Rosch and her colleagues viewed the human organism as part of a highly structured world, with innate perceptual sensitivity to its hierarchical structures of correlated feature bundles (Rosch & Mervis, 1975; Rosch, Mervis, Gray, Johnson, & Boyes-Braem, 1976).

Exemplar and prototype concepts claim to be internal representations of the external environment. Goodman (1978), who is no supporter of similarity-based categorization, provides an example showing why a similarity-based concept cannot be an exact representation of external reality. In his example, Mrs. Tricias went to a textile shop and looked at a book of material samples. She chose a sample, and ordered enough material to cover her sofa and chair, requesting that it be exactly the same as the sample she had chosen in the book. When the material was delivered, Mrs. Tricias was surprised to find that it was cut into hundreds of squares exactly the same size as the chosen sample.

The moral of the story is that a sample is a sample of some of its properties, but not others. In this case, the proprietor had given too much weight to the size and shape of the sample. Goodman (1978) is making the point that internal representations of external reality cannot be *exactly* the same. Yet people do *all* focus on certain specific features, out of the many possible features contained in an object, even though which features these should be is not always specified. Mrs. Tricias gave an order which was not literally exact, yet expected her interior decorator to understand what she meant, and this reader at least, felt surprise at the textile shop proprietor's mistake. Yet a prototype view of concepts does not specify *how* to identify the salient and important features, it is simply assumed that people can do so. Of course, the advantage of attributing identification of salient and important features to the organism's innate tendency to perceive similarity structures means that choice of the features relevant to the comparison between object and category need not be explained. Thus, the identification of salient and important features from a seemingly chaotic confusion is ascribed to the innate tendency to perceive `environmental structure.

1.02.4 Concepts as explanatory beliefs

Until this point, the approaches to meaning describe concepts as formal or structural representations of external reality (whatever that might be). An information or content-based approach to meaning sees concepts as the building blocks of knowledge (Medin, 1989; Rumelhart, 1980). Such a view does away with any assumption about a fixed or truthful reality waiting to be "discovered". Borges' fantasy example of ancient Chinese classifications and Dixon's linguistic classes of the Dyirbal aborigines highlight the difference between the approaches (Dougherty, 1978).

The knowledge-based view does not say that logical concepts and similarity-based theories are wrong, just that they are incomplete (Medin & Wattenmaker, 1987; Murphy & Medin, 1985). An internal representation exemplifies only *some* of the properties of an object in the outside world, and the relevance of the properties can vary with the person's purposes and goals, although Mrs. Tricias's interior decorator does not seem to have realized this. A similarity-based concept cannot account for *the choice* people have to make about which features to focus upon, during a similarity comparison. It has been suggested, however, that physical similarity might be reflecting some nonarbitrary theoretical construct which guides the choices people make about *which features* are the relevant ones during comparison of an object's appearance with the internal prototype (Ruse, 1969; Shanon, 1988b). A view of concepts as explanatory beliefs explains *why* we know which properties to choose as relevant in a particular context or with a specific goal.

All explanation-based approaches emphasize that the effects to be observed in the use of concepts (for example, typicality effects, permissible inferences) cannot be explained by simply considering concepts in isolation. Such effects often emerge through an interaction between the information specific to a particular concept, and background knowledge about the world (Dougherty, 1978; Lakoff, 1987). People tend to construct explanatory concepts to explain why objects, creatures, events et cetera should appear similar, using this background knowledge about how the world works. This background knowledge "grounds" or "embeds" the constructed concept, giving it stability by relating it to the rest of the person's conceptual knowledge (Medin & Ortony, 1989; Wattenmaker, Nakamura, & Medin, 1988).

Murphy and Medin (1985) give two examples of the use of background knowledge in the construction of explanatory concepts, homeopathy and contagion, which are naive belief systems about how objects and people inter-relate in the world. These belief systems seem to recur cross-culturally (Frazer, 1959), so that whilst the *content* of the belief might differ across cultures, the *form* of reasoning it employs remains the same.

Contagion is the principle that a cause must have some form of contact to transmit its effect. In other words, the closer two events are in time and space, the more likely they will be perceived as causally related (Dickinson, Shanks, & Evenden, 1984). Contagion reasoning occurs cross-culturally because of the human propensity to draw causal links even though only associative links might exist (Michotte, 1963). One modern example, found in so-called advanced cultures, lies with the horror some people have of even minimal contact with anyone suffering from AIDS. This is the widely reported case of Yvonne ---, the child who contracted the disease through blood transfusion. She was banned from attending her pre-school centre after the frightened parents of other children lobbied to have her removed. Subsequently, her family had to emigrate to New Zealand, where she found a primary school which would accept her.

With homeopathy, causal links are drawn between two events or objects, even though the link might be one of resemblance only. The physical appearance of the symptoms (or effects) suggests a similar cause, and the medicine prescribed as a cure will often resemble the symptoms (Medin, 1989). In the Azande culture, the cure for ringworm is to apply fowl's excrement because it looks like the ringworm (Frazer, 1959). The claim is that resemblance is a fundamental conceptual tool of everyday thinking in all cultures, not just so-called primitive cultures (Schweder, 1977). Another effect of perceptual similarity upon people's reasoning is the belief that causes and effects should be similar in magnitude. The germ theory of disease was not easily accepted by lay people when first introduced, because it was difficult to imagine how such invisible organisms might result in death (Einhorn & Hogarth, 1986).

These two kinds of reasoning illustrate people's tendency to draw causal relations about the objects and events in the world around them, and as such, lend support to the explanation-based theories of representation. The role of these principles is not so much to constrain similarity, as to use similarity as a guide to a causal explanation. Immunization might be seen as an example of homeopathy where the cure and the disease are very similar. Homeopathy is not the explanation itself, but it acts as a guide to the underlying scientific principle which explains immunization's efficacy as a cure. The principles of homeopathy and contagion are not constraints upon similarity, but rather they act as constraints upon people's search for causal explanations (Medin, 1989).

So the third answer to the question posed at the beginning of this chapter is that our concepts arise from some domain of knowledge, and consist of beliefs about the world. The word *believe* is critical insofar that it broadens the context for the meaning of "concepts" to the psychological and cognitive sense, not only its metaphysical or naive realistic sense (Smith, 1990). This view, that concepts derive their meaning from being embedded in a network of theories is a common one in the philosophy of science (Brewer, 1993; Hempel, 1966; Suppe, 1977). One definition of a theory is as "a set of causal relations that collectively generate or explain the phenomena in a domain' (Murphy, 1993a, page 177).

1.03 EVALUATION OF VIEWS

Ultimately, the conceptual views described above are all concerned with the question of how concepts arise. In other words, does meaning lie within ourselves, or should we look for it in the outside world, its underlying forms or similarity structures? As can be seen from the preceding section, concepts can be viewed as metaphysical phenomena (organizing principles of some objective reality); or approached in an epistemological fashion, as products of the human cognitive system, in the form of representations of knowledge and beliefs.

The theoretical advantage of assuming that the source of conceptual meaning lies in the outside world is that the issue about the stability of concepts is resolved automatically. Concepts are said to be stable because of the world's structures in which they are embedded. However, the question must arise if it is assumed that external reality is not the sole source of meaning: how to account for the stability of concepts?

Rey (1983) has argued in his paper "Concepts and stereotypes" that a primary function of concepts is to provide our world with stability. To achieve stability, a concept must satisfy two necessary conditions: intrapersonal stability and interpersonal stability. He claims that the same type of cognitive state can occur at different times *within* a person, and this intrapersonal stability of concepts provides the basis for the person's conceptual competence. Furthermore, different people can be in the same type of state at the same and different times, and this interpersonal stability provides the basis for comparisons of cognitive states *across* people.

1.03.1 Stability through ultimate Truth

Theories of concepts based upon the metaphysical approach, claim that concepts are stable because they are grounded in the structure of reality. Theories based upon assumptions of metaphysical realism are the most deterministic of all theories of concepts. They take an objectivist view of human cognition, with language and mind being totally dependent upon some metaphysical and truthful structure of the external world. The implication of this view leads to a rejection of the assumption that concepts for natural kinds mediate between words and referents, an assumption at the core of current psychological views of concepts (Komatsu, 1992).

On the basis of such a view, one could argue that concepts are not situated at the cognitive level at all. In order to study them, one needs to map the structure of material reality. There are a number of cognitive theories which posit *a priori* assumptions about the structure of the external world, and then go on to describe the conceptual organizing principles which underlie it. The organizing principles can include mathematical formulas such as Boolean functions or Bayesian probabilities which compute properties (Anderson, 1991b); or logical algorithms which compute mental models of objects, and the locative relations which hold amongst them (Johnson-Laird, 1983); or singly necessary and jointly sufficient defining features (Lyons, 1977a, 1977b).

What Rey (1983) was really saying was that concepts are stable both within persons and across persons because they consist of objective Truth (for example, mental calculations of Bayesian probabilities). But as Kripke (1972) and Putnam (1975b) had pointed out, people need not be aware of such metaphysical truths, or they might even have mistaken beliefs about them. Thus, the information in the mental representation labelled by a natural kind term is not necessarily true of the real-world referents.

Putnam (1975b) argues that one indication that objective truths exist independently of human cognition is that people are so ready to *revise* their concepts. Scientific experts might make new discoveries or change old theories about the world around us, such as discovering that the Earth is round, and people consequently *revise* their opinions. As a result, the meaning of some words cannot be completely captured by an explication of the stereotypes which they label; these meanings are metaphysical truths which exist independently of human cognition. Although he later revised his opinion, Putnam declared in this article that "meanings just ain't in the head" (Putnam, 1975b, p. 227).

The nub of the problem is how would such ultimate truths, which exist independently of human cognition, provide stability of concepts in a psychological sense? Objective, metaphysical truth might exist, but that is not enough, because in order for such truths to have any effect on the stability of a concept, there has to be an awareness of the concept's truth. The earth was round long before most people believed it; and although it is true, it cannot be said that these people actually knew it (Hospers, 1990). So stable information contained in the conceptual core is more likely to be constituted of *beliefs about* what is true, rather than true metaphysical knowledge which can never be proved one way or another.

Hospers (1990) describes such beliefs as the *subjective* component of knowledge, and points out that such subjective beliefs are not enough; a final condition for stability is that there must be good evidence for believing something. This last requirement is the most troublesome of all, hinging as it does on how much evidence is required before one can believe something to be true. So the more evidence there is to reinforce the subjective belief, the less likely a person will be to revise a concept about, say, the flatness of the Earth or the essence of a *tiger* or *lemon*. In short, the stability of a concept can be measured by people's readiness to revise it, and this is easier to do, the more experiential (subjective) evidence there is to support the revision.

1.03.2 Stability through similarity structures of appearance

An alternative account for conceptual stability involves similarity-based concepts, which are based upon assumptions of naive realism. Naive realism is less deterministic than metaphysical realism, but its theories also involve environmental constraints, in this case, physical appearance. Unlike the theories based upon metaphysical realism, the links between the human conceptual system and the objects "out there" remain unspecified. There are no detailed *a priori* assumptions to be made about the underlying structures of the outside world. Theories based upon naive realism usually take a normative or cultural view of the kind of knowledge contained in concepts. In other words, they are taking an epistemological approach to concepts (Gardner, 1987; Lakoff, 1987).

Goodman (1972), however, has nothing good to say about similarity of appearance as a source of conceptual stability, arguing that physical appearance is a straw man:

> Similarity, I submit, is insidious.... is a pretender, an impostor, 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, p. 437).

Goodman argued that a general notion of similarity cannot give even a partial account of conceptual stability, because always it is conceived of as being relative to something else ... "with respect to". Consequently, to define similarity as does Tversky's model (1977), that "*a* is more similar to *b* than it is to *c*" is a meaningless notion, unless one can also specify *in what way* they are more similar, thus giving the comparison a frame of reference and specific meaning.

One example of the need for a frame of reference (other than similarity of appearance) is found in the difficulty of imperfect community, and cited by

Goodman (1977). This difficulty undermines both set-theoretical models which postulate defining features of physical similarity, and also prototype theories of similarity. One example of imperfect community comes from the *Marsupials* category which lists such members as *kangaroos, opossums* and *marsupial mice.* These last are more similar to ordinary *mice* than they are to *kangaroos,* yet (zoologically) they do not belong in the same category as *mice.* Set-theory would predict that no one thing remains outside the category-set which resembles the member of the set even a little, that the category-set is a "perfect community". Prototype theory would predict that *marsupial mice* belong in the same category with ordinary *mice.* It would seem that similarity cannot be the "glue" which coheres members into a category after all.

Quine (1977), who provided the *Marsupials* example, also has an explanation for it. Unlike Goodman, Quine (1977) believed similarity to be fundamental to human thinking and reasoning. He explained the paradox/difficulty of imperfect community by drawing a distinction between *subjective* and *objective* similarity metrics. The two metrics can be distinguished by saying that subjective similarity is innate, and always "with respect to" something else because it is based upon physical appearance. For example, *a* is more similar to *b* with respect to shape, or colour, or size, or animacy.

Objective similarity tends to look at explanatory relations and so plays a diagnostic (or conceptual) role in categorization. Consequently, *marsupial mice* and *ordinary mice* would not be seen as two of a kind if the categorizer were using a similarity metric other than physical appearance, such as the ordinary mouse's genetic composition or naive theories about what constitutes membership in the *Marsupials* category. On the other hand, subjective similarity might be the highly flexible perceptions of similarity by single individuals, consisting of idiosyncratic (or experiential) knowledge (Quine, 1977). It would seem that physical similarity is indeed "with respect to" a frame of reference, and as such, it provides an insufficient grounding for conceptual stability. Thus the same conclusion is arrived at here, as was concluded in the last section: that individual experience does have a role to play in concepts and the knowledge they contain. An *adequate* similarity-based view of concepts would require a *subjective* component, as well as a normative component of knowledge.

1.04 ARGUMENT AND HYPOTHESIS: STABILITY THROUGH THEORY-EMBEDDED CONCEPTS

The first issue discussed in this chapter, which is unresolvable by empirical means, concerns the source of stability for concepts (Hampton & Dubois, 1993). If the source of meaning does not lie in the metaphysical structure of the environment, how can an adequate theory account for the stability of concepts both within people and across people. A number of possible answers were considered. Firstly, it was posited that objective Truth and objective knowledge would ensure the stability of concepts, but it was shown that this kind of knowledge is probably unknowable. Secondly, the structure of appearance is posited as a source of stability, but it was shown that perceived similarity tends to lead to instability since people do not always focus upon *exactly* the same features, so that anything might be similar to anything else. It is argued that subjective similarity might be part of the answer, because it does provide within-person stability of concepts, through self-reference and experiential evidence. Finally, general knowledge and beliefs about the world around us might prove to be a complete source of stability for concepts, because this approach makes use of both normative and experiential knowledge.

Rey (1983) drew a distinction between the metaphysical and epistemological approaches to concepts. What is true in any possible world, but upon an "unknowable" metaphysical plane, need not be the same as what is conceivable and compatible with all we know (epistemologically possible). This distinction leads to Rey's (1983) distinction between perceptual features of physical appearance (epistemology), and diagnostic core features (metaphysics).

Especially where "natural kind" concepts are concerned, like those of biology, these two kinds of features do not always coincide. Something might share all the usual perceptual features of a kind of plant or animal, yet fail to *be* that kind: silk flowers, wax figures such as found in Madame Tussaud's, and fancy mechanical toys. It would seem that the distinction between outward appearance and essential reality is psychologically real for most people. If this were not the case, then any number of people might use different features when deciding whether a creature is a *Bird* (not a *Fish).* The result of different people using different features to identify an animal or a plant would be complete instability of concepts. Rey's (1983; 1985) central thesis was that what accounts for inter-conceptual stability and identification (metaphysical truths) cannot account for categorization (lay stereotypes). He is saying that the categorization and identification of objects are not the same thing. Our representations of lay stereotypes may be open to revision, but not metaphysical truths.

On the other hand, Smith, Medin and Rips (1984) argue that, as long as Rey (1983) equates stability with the *sameness of concept*, he probably is correct. Sharing the *same concept* across people and within one's own beliefs and preferences, requires a specification of identity conditions. But the researchers suggest another sense of stability, which can be equated with *similarity of mental contents* (between persons stability). In this sense, what accounts for categorization can and will at least partially account for identity and stability. To emphasize the "sharing of mental contents", they call this interpersonal stability "communality". Whilst no philosophical arguments can justify communality, there is plenty of empirical evidence for it, much of which was provided by Rosch and her colleagues.

Psychological studies of concepts must necessarily take an epistemological approach to them, that is, look at the knowledge or information they contain (Smith, Medin & Rips, 1984). As was explained in section 1.02, each view advocates a different kind of knowledge of the world as being contained in a concept. It is suggested here that the actual information content of a concept might be of some influence in the stability of the concept, to what degree it will be revisable or subject to change This leads to the second issue which was addressed in this chapter. The study of metaphysical truths as a basis for concepts can only be speculated upon and not empirically resolved. If the study of concepts is restricted to their informational contents used during categorization, there can still be an accounting for both kinds of stability in concepts.

Quine's (1977) distinction between objective and subjective similarity is useful here.It implies that objective similarity might be the constructed beliefs consisting of normative (or social/cultural) knowledge, thus providing stability of concepts across people. A valid account of concepts and their informational contents would incorporate both components of knowledge, normative and subjective. It is suggested here that only a theory-embedded view includes this subjective component of knowledge which provides concepts with their stability within a person, so that a person is likely to use concepts in a consistent fashion, at different times. It would also provide the *experiential evidence* at an emotional level, which reinforces beliefs and which Hospers (1990) said was so necessary for concepts to resist revisability Only a theory-embedded view can account for the fact that people not only know by what features to identify a creature in a number of different contexts, but they also choose the same, or similar, ones as other people do. The view of concepts as building blocks of knowledge can account for the *choices* people make as to what constitutes relevant information during categorization decisions and concept formation.

In summary, theories which can account for the interpersonal stability of concepts (concepts which are communally shared) must also account for the intrapersonal stability of concepts (consistency of concepts used by the one person). Thus, dictionary or normative information is not enough because it does not take into account the individual, and his or her construal or understanding of personal experiences. Knowledge constructed from personal experience will ensure within-person stability for concepts, because self-reference is what has the most relevance for individual people.

It is argued here that a concept's experiential details and information (such as individual goals, needs, explanatory context) will contribute towards entrenching or "grounding" a concept, thus making it more stable than one based upon purely physical similarity, dictionary definitions or scientific theories about unknown/unseen essences. One hypothesis tested in this thesis is that, whilst normative information is necessary for a person to function as part of a community, idiosyncratic information also has a part to play in the stability of concepts. If so, it is suggested that people will be likely to find information concerning personal needs and goals to be of the utmost relevance during categorization, whether it be the needs and goals of themselves or of other people.

CHAPTER TWO

REVIEW OF RESEARCH INTO CONCEPTS AND CATEGORIES

2.00 CHAPTER OVERVIEW

Section 2.01 introduces three theories about categories: the classical, prototype and explanation-based models. The two properties of coherence and flexibility, which an adequate theory needs to account for, are described. In the next three sections, each theory is outlined in more detail, with a description of how people categorize. Examples are then presented, which run counter to the particular theories. The goal is to weigh up the various inadequacies of each theory.

Section 2.02 looks at the early research in support of classical models, such as the logical hierarchy of Collins and Quillian (1969), and the subsequent empirical evidence against it. An updated version of the classical approach, Anderson's (1991a) theory of adaptive categorization, is outlined. The classical approach is evaluated with regard to how well it fulfills the categorization principle of cognitive economy.

Section 2.03 deals with the major features of prototype theories. There are four variants of prototype theory: as abstract amalgamation, as independent feature lists, as specific exemplars; and as correlated bundles of characteristic features. These variants are evaluated as to how adequate their constraints upon categorization might be. It is claimed that they cannot account for two properties found in most categories: coherence of items as a category and flexibility of relations between concepts.

Section 2.04 describes some explanation-based models, including the two-stage model of categorization (Smith, Shoben & Rips, 1974), and Rosch's (1983) theory of dual representation. The "transformation" studies of Carey (1985), Keil (1989), and Rips (1989),which provide empirical support for a dissociation between similarity and categorization judgments, are described.

In Section 2.05, it is argued that, of all three approaches, the last one (which includes explanation-based accounts of categorization) is the most psychologically real one. It is this approach's category models which incorporate the two properties of coherence and flexibility in their accounts of people's category behaviour,. The models achieve this by specifying what constitutes information *relevant* to the concept and its category members. It is proposed that: (a) the representation, structure and process of classical and prototype models is inadequate; and (b) that the most relevant information contained in concepts will be concerned with personal details, such as goals, needs, purposes. Finally, section 2.06 describes three empirical studies which aim to test such arguments.

2.01 ISSUES AND AIMS OF CHAPTER

Chapter 1 considered the question of why certain concepts arise, and described metaphysical and epistemological approaches to the question. Their main area of difference lies in their explanations for the stability of concepts: the former claims that metaphysical truths underlie conceptual stability, and the latter claims knowledge about the world ensures stability. It was suggested that, since people might not know objective (or "truthful") knowledge, a metaphysical approach cannot be empirically tested in psychological studies. Further, conceptual stability between people might be based upon normative information, but it is also suggested that subjective information might come closer to capturing conceptual stability within people. Chapter 1 concluded in favour of the view that concepts are stable because they are embedded in beliefs and theories about the world.

Chapter 2 looks at the implication of this answer with regard to categories. One approach taken by past researchers has been to study what kind of concept makes a category cohesive (Medin, 1989; Medin & Smith, 1984; Murphy & Medin, 1985). The question of why we have the categories we do (and not others) is addressed in this chapter, and the empirical research which has investigated it in the past is described and evaluated.

Murphy and Medin (1985), in their article on the role of theories in conceptual coherence, asked why a given set of objects grouped together should form a category which is sensible, informative, and useful; whereas another grouping of items might seem vague, absurd or useless. As an example, they propound an old Biblical puzzle which lists the abominations of Leviticus, upon which the Jewish dietary laws are based, and from which the categories *clean* and *unclean animals* can be derived.

> Why should camels, ostriches, crocodiles, mice, sharks, and eels be declared unclean, whereas gazelles, frogs, most fish, grasshoppers, and some locusts be clean? What could chameleons, moles and crocodiles have in common that they should be listed together? That is, what is there about clean and unclean animals that makes these categories sensible or coherent? (Murphy & Medin, 1985, p. 289)

The research literature has produced three main approaches to people's category behaviour, and each would treat the Biblical puzzle differently, giving separate accounts of why this given set of objects should group together as *clean* or *unclean animals.* For example, a classical approach would argue that the animals possess defining features which make them either clean or unclean.

Anderson's (1991a) theory of adaptive categorization is an updated version of the classical approach, and it would describe the groups of *clean* and *unclean animals* as classes rather than "true" categories, because they are not derived from the objective structure of the environment. Anderson (1990) dismisses the two groups as an example of confusion between categorization and labelling. The dissimilar creatures listed (for example, camels, ostriches, crocodiles, mice and eels) come from different natural categories, but they do share a label *unclean animal* which places them in the same class (that is, a nominal category). Anderson's (1990) view of the two categories in the Biblical puzzle, then, would be that membership is determined by the definition contained in the category-label, as might be found in any dictionary. This is not satisfactory, as the account fails to specify what the definitions for *clean /unclean* might be.

A prototype approach might say that they are really the one category, *Animals,* with each instance having a different attribute value on the dimension of *cleanliness.* This would be a structural description of how the different *Animals* differ in their degree of *cleanliness,* in how good an example one animal might provide of an *unclean animal* or otherwise. It does not explain where the cut-off point, or boundary, between *clean* or *unclean* might lie or how to decide upon it.

Murphy and Medin's (1985) paper advocates a theory-based approach to categorization, and makes the point that people will form coherent categories of objects on the basis of their beliefs and knowledge about the world. Their theory addresses the fact that to solve the puzzle, one needs to know what is believed to constitute cleanliness or uncleanliness in an animal. Douglas (1966) has suggested a possible belief for this particular example, which holds explanatory relations between type of habitat, biological structure, and form of locomotion. Those creatures which are not biologically equipped for the right kind of locomotion in their element would be considered unclean. Thus, creatures which live in water should have fins and scales, and swim; creatures which live on land should have four legs and jump or walk; and creatures of the air should fly with feathered wings. Hence, penguins would be unclean because they do not fly. Such an explanation would point to a partitioning of the category *Animals* according to underlying conceptual principles.

The two main issues raised by this example concern the coherence of categories and the flexibility of between-concept relations. Firstly, the example violated Western notions of what constitutes "naturalness", or as Smith (1990, p. 34) would describe it, what "seems to belong together" coherently. Yet, once the underlying religious belief was explained to the reader, the two categories of *clean* and *unclean animals* were comprehensible and coherent as such. Thus, the first issue considered in this chapter is what the various empirical studies have to say about the *coherence of categories.* Secondly, concerning the example, the reason for the sudden comprehension was the shifted focus provided by the explanation. When the relevant features which constitute the attribute of *cleanliness* and the attribute of *uncleanliness* in an animal were made salient, the various creatures cohered as either one or the other category. This second issue concerns the *flexibility of concepts*

Coherence is what happens when certain groupings of objects seem to naturally "hang together" or to "make sense" (Goodman, 1972; Osherson, 1978). As a principle, world structure has been used by the classical and prototype theories to explain the coherence of items into categories. *A priori* assumptions are made by these theories, either about the underlying principles which organize world structure, or about the perceived appearance of world structure. Categories are then said to "naturally" hang together or cohere (Komatsu, 1992). However, the coherence of categories which are not biologically natural (such as ad hoc category-types) cannot be explained by world structure.

Flexibility is what happens when more than one category can be found for a group of objects. As a principle, cognitive economy is desirable because it aims for a minimum of effort in storage and processing of concepts (Anderson, 1991a; Rosch, 1978). However, one purpose of concepts is to access knowledge about the category which will allow the categorizer to predict events or make inferences about individual members, so that maximum storage of information is also desirable (Medin, 1989). Consequently, an even balance between the two objectives of cognitive economy and informativeness needs to be maintained by any *adequate* theory (Komatsu, 1992).

The general aim in this chapter is to review the various psychological theories on what it takes to form a comprehensible category, using each theory's incorporation of the properties of coherence and flexibility into its view of categories, to evaluate its adequacy. More specifically, Chapter Two aims to show that the more recent research supports a theory-based view of categorization. These studies find their participants using conceptual knowledge during categorization; and their perceptions of similarity are found to be more flexible than older and more traditional theories had assumed.

2.02 CLASSICAL MODELS

An intuitive idea about the nature of categorization which has held sway since Aristotle is that all members of a category must meet some defining rule which determines membership. Aristotle would have claimed that categories can be defined by unseen essential features such as *an ability to fly* or *a light skeletal construction* (Putnam, 1975a, 1975b).

2.02.1 Logical hierarchies

The tradition (e.g., Bruner, Goodnow & Austin, 1956) of using artificial concepts in category research was broken by Collins and Quillian in 1969. In their studies, they used items from biological classes such as *Birds, Fish* and *Mammals* and everyday categories such as types of *Drink.* These categories were to be defined by sets of individually necessary and collectively sufficient attributes or features. For example, to be classified as a *Bird,* an *Animal* must have certain physically defining features, perhaps *feathers, wings* and *a beak.* If all three features are present in some animal, they are sufficient for it to be recognized as a *Bird.* If one of the three defining features is missing in a potential member, then it fails to fulfil the categorization rule, and does not belong to the category defined by the rule (Katz, 1972; Medin & Smith, 1984).

Collins and Quillian (1969) produced a hierarchical model of logical categorization which was based upon the logical application of necessary defining features. Their hierarchy very much resembles a zoological taxonomy: at the top are broad general superordinates like *Animals;* which in turn break into more specific categories of *Mammals, Birds, Fish;* becoming more specific on levels further down as in *ostrich, eagle, canaries.* At each level in the hierarchy, the categories are linked to a list of necessary properties such as *eats* or *has feathers.* Categories at lower levels link only to special properties which are not generally applicable, for example, *can sing* for *canaries* but not for *eagles;* or *wings* and *feathers* apply only to *Birds,* but not to *Mammals* and *Fish* which are on the same level as *Birds.* The defining features apply to all *Birds,* and are listed at that superordinate level, but not the *eagle* or *ostrich* level, and in this way, cognitive economy of storage is achieved, as features are listed only once.

The processing assumptions are logical in that processing time is said to increase with the more levels the subject has to search through. For example, to verify the statement "A shark is an animal" involves a jump of two levels (*Animal - Fish - shark*) in order to make a decision, whilst "An ostrich is a bird" involves only one level. The former proposition should take longer to verify than the latter. Likewise, statements involving properties specified at higher level categories should take longer to verify than category statements, since the person must retrieve such a property as well as locating the category which possesses it. For example, "An ostrich has feathers" would take longer than "An ostrich is a Bird". The model is precise and elegant, and it might fit an expert's detailed mental representation of the animal kingdom, but lay people just do not have such a logical organization of knowledge. Subsequent studies (listed below) showed that subjects' verification response times could be influenced by three variables which have nothing to do with logic: frequency of association, semantic relatedness, and typicality.

Frequency Effects

Being a logical hierarchy, facts were said to be stored only once, thus fulfilling the principle of cognitive economy. For example, *can fly* would be stored with *Bird* only, not with *robin* and *canary* as well. From this, predictions for response times for verification of property statements were driven by search time through levels of the hierarchy. Conrad (1972), however, found that subjects generated such properties for all levels of the hierarchy. Furthermore, it was the frequency of production of such properties, rather than the hierarchical level at which they should logically be found, which drove speed of response times during verification tasks. Verification for property statements like "A robin has wings" were faster than "A robin has feet". *Wings* is a high-frequency property for the concept *robin* while *feet*, being a low-frequency property, required more time for retrieval (see also Ashcraft, 1978a, 1978b).

Response times for category statements also were influenced by familiarity, rather than a category's place in the levels of the hierarchy. Rips, Shoben, and Smith (1973) showed that response times for category statements, such as "a dog is a mammal", took longer to verify than "a dog is an animal" even though the former involves a search through fewer levels. The most probable reason is that *Mammal* is a more unusual category in an ordinary subject's experience, and that *dog* is more frequently associated with *Animal*. Frequency of association, then, was a better predictor than place in the levels of a logical hierarchy.

Semantic Relatedness

The more related two concepts are in semantic memory, the faster the mental search process which retrieves information about them. Semantic relatedness or similarity effects were found to affect not only positive categorization statements, but negative ones also (Smith, Shoben & Rips, 1974). According to the model, the question "Is a bat a precious stone?" should require a longer time to search before a negative response could be given, than "Is a bat a Bird?", which is only one level away. But Smith et al. (1974) showed that a negative response decision is based upon degree-of-similarity or meaningful relatedness between the two items, not their position in the logical hierarchy. A *bat* has many features similar to a *Bird*, and this seems to create cognitive dissonance, confusing subjects to the degree that they take longer to respond negatively to this proposition, than to the one regarding a *bat* and a *precious stone.*

Typicality Effects

However, the main contradiction to a logical model concerned categories on the same level. The model implies that all instances *within* a given category are of equal status because, in order to be members, it is necessary that they all must contain certain defining features. However, it was found that typical members of the same category can be judged more rapidly than atypical members (Rips, Shoben & Smith, 1973; Rosch, 1973). For example, subjects' more rapid response times for 'a robin is a Bird' than for 'a parrot is a Bird' were found to be correlated with, and predicted by, their typicality ratings. It was this finding in particular which led Eleanor Rosch to begin her studies into the effects of typicality, producing a gradient structure of items internal to the category.

2.02.2 Rational algorithms

Anderson's (1991a) theory also describes category boundaries which are clearly demarcated because, as did Collins and Quillian (1969), he proposes that membership of an item rests on its possession of necessary and sufficient features which define its category-membership. As in the logical hierarchies described above, Anderson (1991a) also describes the environment as being divided up *a priori* into categories consisting of mutually exclusive and disjoint sets of objects. This is an objectivist view of the environment, since he claims that the categories exist independently of human cognition. Such a view places Anderson's (1991a) theory in the category of classical approaches. Environmental structure is said to take the statistical form of Bayes theorem.

In Anderson's (1990) view, thought and cognitive phenomena are governed by two principles - optimality and rationality. Contrary to all expectations, memory functions according to an *optimality* principle, that is, it is exquisitely tuned to the statistics of information presentation in the environment. Where categories are concerned, optimal performance would consist of discovering the one "true" category for an object or living thing, and once this was achieved, drawing the highest number of inferences about the item from its category.

According to Anderson (1991a), the category of an object or creature will be determined by whichever category is likely to be most used or most needed across situations and cultures. This leads to the second principle prescribed by Anderson as underlying the adaptive character of thought, which is that of *rationality*. By this, he does not necessarily mean that categories must be logical, but that they must operate according to whatever information is most needed and used by the individual. Research showed that human memory displays the fastest retrieval latencies and highest probability of recall for the information that is statistically most likely to be needed (Anderson & Milson, 1989). Rationality has been described as an environmental constraint upon memory, in that perception of the statistical cues in the environment is influenced by the need for information and how often that need tends to repeat itself in various situations. The rationality principle could be said to govern availability of recalled information by ensuring that the information which is most often used is the information most often needed.

Anderson (1990) sees the nature of "true" categories as mutually exclusive, disjoint sets of objects, since that is how the environment is divided up; thus the nature of categories is not fuzzy, but well-defined. The human categorization process does not always mirror the environment perfectly, or achieve perfect predictability, but its goal is to achieve disjoint partitions of the objects in the world into mutually exclusive sets. Such partitions should maximize what we can predict about the world. We cannot predict features with greater accuracy than their objective base rate.

An object can belong to only one "true" category at a time and all the others are just linguistic labels which need to be predicted like any other feature. To calculate the probability of an object belonging to a category at each point in time, one needs to have a fixed set of categories, one needs to update these categorical hypotheses as each object comes in, and to do so with a limited amount of effortful computation. An iterative algorithm which satisfies these constraints has been proposed (Fisher, 1987; Lebowitz, 1986; 1987).

An iterative algorithm means that the algorithm has to be incremental, and commit to a hypothesis after every object seen. This contrasts with some algorithms (e.g., Quinlan, 1986) which take in a large number of objects, process them, and only then deliver a set of categorical hypotheses. It is also in contrast to typical clustering algorithms (Anderberg, 1973). The case for the use of an iterative algorithm is the simple fact that people need to be able to make predictions all the time, *not* after seeing many objects and much thought.

In developing his theory of adaptive categorization, Anderson (1991a) took into account the empirical evidence which showed differences in reaction times between members of the same category. He attributed these differences to order effects brought about by the order in which the various items of the category were learned, such orders not being predictable across subjects. Such differences in reaction times would vary randomly across individuals, since the order would be relative to the individual frequency of experiencing the object. Thus, internal ordering of same-category instances (that is, category-structure) would be different for each individual.

This category structure is developed by assigning each incoming object to the category it is most likely to come from. Given the cues in the environment, the categorizer cannot know for sure what to expect. However, the individual begins with weak hypotheses about environmental structure, and possible categories, and with experience makes these increasingly strong. It is this process of updating the individual's probabilistic model of the environment which leads to a Bayesian statistical inference scheme, using a Bayesian decision theory (for example, Berger, 1985). Note, however, that the prediction for a new object is *not* calculated by determining its most likely category, and the probability of its having a certain feature given that category. Rather, a weighted average is calculated *over all categories.* This handles situations where the new object is ambiguous or unclear among multiple categories. It will weight these competing categories approximately equally. The two main assumptions of Anderson's (1990) rational model of categorization - that the environment has a statistical structure and that human cognition is innately adaptable to it - imply that the "categories in the human head" are represented by a fast categorization algorithm which takes the form of a statistical computation along the lines of Bayesian probabilities. The relationship between task and cognitive performance is monotonic and unmediated (Gigerenzer, 1991). Since environment is divided up into a fixed number of ontological categories, their category sizes form the base rates used in a rational Bayesian calculation of probability.

2.02.3 Evaluation

The main problem with Anderson's theory is his metaphysical assumption that the environment is divided up *a priori* into a fixed number of ontological categories, because there is no way of knowing whether this is really so. As a counter-example, Watanabe (1969) offered formal proof (theorem of the Ugly Duckling) that the same creature can be categorized equally logically either as a Beautiful Swan, or as an Ugly Duckling. He showed that no "objective" justification exists for preferring any one partitioning of entities in the world over other possibilities.

The classical models depend overmuch upon the principle of cognitive economy in formulating their description of membership criteria, so that the individual members of the extension are impoverished in terms of their informativeness. If the model is Aristotelian, the shared criteria can be logically defining or essential features of the same ontological category (Collins & Quillian, 1969; Putnam, 1975a, 1975b). If the model is Platonic, the criteria consist of shared abstract forms, such as Bayesian algorithms, which are innate tendencies (Anderson, 1991a). Either way, the conceptual criterion which determines category membership contains no structural information about the individual members, and so cannot account for predictable typicality effects across individuals.

The classical approach to concepts described by hypothesis-testing, logical hierarchies and computational algorithms completely fulfils the principle of cognitive economy in memory storage and processing, because the sole representation is a membership rule which categorises members. The disadvantage is that unclear cases are left unaccounted for. For example, the *Pope* is technically a *bachelor* or *unmarried male*, but how many people would categorize him thus? Would everyone categorize *carpet, clock* or *radio* as pieces of *Furniture*? Such unclear cases mean that people disagree with each other, and contradict themselves across separate occasions, about category membership of very atypical items (Barsalou, 1989; Bellezza, 1984a, 1984b; McCloskey & Glucksberg, 1978; Medin, 1989).

Further, classical category models fulfil the cognitive economy principle at the expense of discarding too much information, such as information about frequency, semantic relatedness and typicality (Komatsu, 1992). They are omitting potentially vital distinguishing characteristics of exemplars of the same category. The result is rigidity of categorization. For example, the fact that a *cat* belongs to the *Feline* category is stored directly and economically, but no account is given of how the *cat* can be differentiated from a *tiger*. Thus, since only the membership rule is represented, access to knowledge about a tiger's danger potential is not theoretically possible.

In summary, then, classical models of categorization cannot accurately specify defining features, explain unclear cases, nor account for effects of association frequency, semantic relatedness or typicality. Their account of conceptual coherence is based upon unknowable objective knowledge, and their *a priori* assumptions about world structure create the problem of rigidity of categorization.

2.03 FUZZY PROTOTYPE MODELS

Whilst Anderson's (1990) recent revival of a purely classical approach is elegant and parsimonious, it leaves a number of questions unanswered satisfactorily, such as the predictability of typicality structures across subjects, and the use of underlying conceptual principles during categorization. Both theoretical and empirical factors have contributed to the downfall of the purely classical approach to categorization. Theoretically, although it might seem intuitive that categories be represented solely by a definitive rule, philosophers and linguists have failed to agree upon the defining features necessary for a lexical concept (McNamara & Sternberg, 1983; Medin, 1989). Wittgenstein (1953) identified the problem as follows.

> Consider for example the proceedings we call 'games'. I mean board-games, card-games, ball-games, Olympic-games, and so on. What is common to them all? Don't say 'There must be something in common or they would not be called games', but *look* and *see* whether there is anything common to *all*. For if you look at them, you will not see something that is common to them *all*, but similarities, relationships, and a whole series of them at that... I can think of no better expression to characterise these similarities than 'family resemblances'. (*Philosophical Investigations*, 1953, pp.31-32)

Empirically, Rosch overcame the difficulty of pinpointing necessary and sufficient features which were common to all members of a category by studying their characteristic features instead. Her studies were based upon stimulus-items which consisted mainly of the names for natural categories like *Fruit* or artifactual (man-made) categories like *Weapons*. Further studies have also been conducted on other category-types including perceptual ones of form or colours (Rosch, 1973), artificial stick figures, dot patterns, letter strings (Rosch, Simpson & Miller, 1976), superordinate semantic categories (Rosch & Mervis, 1975), and lexical and picture categories of living things and artifacts (Rosch, 1973, 1975a). The overall result was that categories possess an internal graded structure, where some members are more typical (better examples) of their category than others.

Rosch's studies provided the main bulk of the empirical evidence against classical categories, and in favour of fuzzy representations of categories. The presence of predictable graded structure in animal and artifact categories contradicted the prediction made by a classical approach - that all instances of a category have equal membership status. Instead, results showed that members of all category-types differed in their degree of goodness as an example of their category, bringing Rosch to the conclusion that typicality is fundamental to people's mental representation of categories. Since then, graded structure has been found to be present in category-types other than natural categories (Barsalou, 1985, 1987; Cohen & Murphy, 1984; Homa, 1984; Oden, 1977; 1987; Rosch & Mervis, 1975; Smith & Medin, 1981).

The main points of prototype theory are described below (Eysenck & Keane, 1990). This is followed by a discussion of four variants of the prototype approach to concepts and categories.

(1) The classical theory requires an absolute rule or condition which must be met for membership in the category to occur. In contrast, according to prototype theory, the representation of the concept can take the form of a prototype, which requires either a composite of characteristic features (abstract or concrete), or the best example (or small set of examples) of the concept.

(2) Although prototype theory accepts that necessary features may be present, as in classical models, they are not jointly sufficient. Membership in a category depends upon characteristic (rather than defining) features, which are considered more typical or representative of the category than others.

(3) Hence, in prototype theory, it is implied that, because what actually constitutes a member is not defined, boundaries between categories can be fuzzy and unclear. Consequently, some members of a category might just as equally be considered members of another category (for example, *tomatoes* as *Fruit or Vegetables),* thus allowing prototypes more flexibility than do classical models.

(4) The result is gradient structure, where instances of a concept can be ranged according to their typicality. This typicality gradient is said to capture the differing degrees of membership of examples of the concept.

(5) Category membership is determined by the similarity of an object's attributes to the category's prototype, whether that prototype be represented by characteristic features or an exemplar of the category.

Over the years, Rosch and her colleagues have proposed at least four different descriptions of what constitutes a prototype model, which would account for their empirical results (see Appendix C for brief summaries of studies carried out by Rosch and her colleagues).

Gleitman, Armstrong & Gleitman (1983) described the family resemblance principle as an analogy to the Smith Brothers, as depicted on the wrappers of popular cough drops. In any family some members share some features in common (e.g., blue eyes), while other members share others (e.g., blonde hair). The closer members share more features in common, but no two members of the family (except identical twins) have an identical set of features. In the same way, typical category members are those with a high degree of family resemblance: they share many attributes in common with other members of the same category.

The four variants differ mainly in their account of how similar features are computed Both family resemblance and exemplar models require more complex representations than do classical models (Smith & Medin, 1981). In the following sections, both the next two variants of prototype theory are based upon the family resemblance principle, differing mainly in the rule used to compare the attributes. The third variant differs from the first two in that it describes the conceptual representation as a specific instance or group of specific instances, rather than representations of general category descriptions. The fourth variant proposes *correlated* feature bundles, rather than independent feature lists, and so is including more information in its prototype than any of the others.

2.03.1 The prototype as an abstract amalgamation

This prototype is a kind of composite, in abstract form, of the most typical members of the category. At this time, Rosch considered the prototype to be a kind of composite, or amalgamation in abstract form, of the clearest cases of the category, constituting a "core meaning" (1973,p.140). It was similar to a "mental image" (1973,p.142) composed of these clearest cases, but did not necessarily have to be an image. The central tendency of an abstract prototype would be based upon a statistical *median*. For example, according to this definition, the representation of the category *Fruit* should be a mental abstraction of the best examples of *Fruit:* say *an orange, an apple and a banana.*

Internal graded structure in a category's internal membership had been conclusively demonstrated, and Rosch attributed it to a fuzzy prototype which, however, there was some difficulty in pinning down. Her 1975a study had indicated that subjects responded more readily to pictures than words, suggesting that a mental image might be closer to the nature of the underlying representation than words. Instances differed in the degree to which they "fit" the prototype and so created a graded membership structure. She concluded that the members are arranged according to degree of typicality, with the most typical members found at the "centre" of a dimensional space, and close to the category prototype, while the least typical ones are relegated to the edge or "periphery" near the boundaries of a postulated dimension of typicality (Rosch, 1973, page 112). This spatial dimension consisted of continuous values along dimensions such as size, colour, shape. She emphasized that the typicality dimension was an abstraction, as was the prototype.

2.03.2 The prototype as an independent feature list

An alternative solution to the difficulty of defining "prototypes" was that of characteristic, independent feature lists put forward in Rosch and Mervis (1975) and is the traditional notion of "family resemblance" derived from Wittgenstein (1953). Rosch and Mervis (1975) broadened Reed's (1972) definition of the distance model (where the averaging of attributes defined the prototype) to include the principle of family resemblance. This principle stated that members of a category needed to share one or several similar attributes with one another and/or with the category prototype. In the feature list model of Rosch and Mervis (1975), the central tendency or distribution of members would be analogous to the *statistical mean.* The set of discrete feature values in a prototype constitutes a summary description for the category (Medin, 1983). The concept *Orange* would be a summary representation such as a list of characteristic features. The concept abstracts across specific instances of *oranges* to give information about what oranges, on the average, are like.

Both these variants are probabilistic in that they require a set of characteristic attributes or features, different weightings which reflect different degrees of importance of each attribute to the category; and some "weighted attribute combination rule" to determine category membership, such as the calculations of the median or mean (Barr & Caplan,1987; Hampton,1979; Posner & Keele,1968; Rosch,1978). Representativeness is said to be based upon the "weighting" of shared attributes according to their importance for conferring family resemblance to the category. That weighting is the function of the number of category instances (and noninstances) that share the attribute (Komatsu, 1992).

"Weighting" or cue validity has been defined as the conditional probability that an object is in a category, given that it has some cue (or attribute) associated with the category. When an object has many attributes associated with a category, then it has high cue validity, whereas a poor category has only inconsistent cues. Thus attribute-matching between the features of an object in the world and one of the family resemblance prototypes would coincide with the notion of cue validity, because the attributes most frequently distributed among members of a category and least frequently distributed among members of contrasting categories are, by definition, the most valid cues to membership. Thus, Rosch and Mervis (1975) pointed out that the principle of family resemblance could mean both features common to members of a category (cue validity) or formal criteria for category membership (abstract prototypes).

To summarize, whether family resemblance is based upon the continuous values of a dimension or the physical discrete features of individual members, the object referents were considered to be prototypical of their category as a whole to the extent that their features overlapped with the prototype's attributes.

2.03.3 The prototype as a specific exemplar

With exemplar models, the prototype representation of the *Furniture* category might be its *chair* exemplar or a set of specific exemplars (like *chair, table, bed).* According to the exemplar view, the representation for the category *Vegetables* is not a single mental object somewhere between a *potato, carrot and peas* but a number of mental objects, each corresponding individually to one of these instances, which have been personally experienced at one time or another. The prototype's central tendency (or distribution of members) would be analogous to the statistical *mode.*

Here an object is a member of a category to the extent that it is close to these best examples of the concept (Rosch, 1975b). It is the frequency of an object's occurrence which determines how it is represented. Exemplar models require a frequency-counter which estimates how many times the various instances of a category have occurred, in relation to other instances. This determines the instance's typicality of its category. Consequently, the prototype is represented in terms of the best member (or small set of best members of the category) which the individual has encountered (for example, Brooks, 1978; Hintzman & Ludlam, 1980; Medin & Schaffer, 1978).

Representative members become salient points in a domain, and the category tends to form around them so that they become representative of it. However, the "distance" between any two concepts need not be symmetrical. A zebra may be judged to be more similar to a horse, than vice versa. Rosch (1975b) uses this phenomenon to motivate her theory that salient members such as *horse* act as cognitive reference points. For example, "A *horse* is essentially a *zebra*" does not ring as true as "A *zebra* is essentially a *horse*", indicating that *horse* acts as a better cognitive reference point than *zebra*. In her 1975b paper, "Cognitive reference points", Rosch suggested that the most typical instance of a category might act as an ideal-type anchor to which other instances are seen to relate. She was not convinced by this theoretical interpretation of individual exemplars as prototypes, because specific prototypes would not include all the relevant shared properties of a category, nor would they be economical in terms of cognitive storage.

Many category researchers subsequent to Rosch, however, have not shared her reservations about exemplars. They have produced a number of exemplar models of categorization, and these include the instance model of Hintzman (1986); the analogical model of Brooks (1978; 1987); and the indirect categorization model of Smith (1978). The result has been no clear consensus on what an exemplar representation should be. Unlike the family resemblance view of attributes, the attributes which hold true for one exemplar need not hold true for another instance. Thus a concept could be a number of representations, with individual ones corresponding to a different exemplar of the concept.

Generally, the exemplar models assume that people evaluate the similarity of a new item to representations associated with alternative categories (Hintzman, 1986; Medin & Schaffer, 1978; Nosofsky, 1988, 1989; Reed, 1972).

For example, Medin and Schaffer (1978) have proposed a context model that assumes categorization depends upon stored information about category members rather than on overall category information. Instances are said to be categorized on the basis of their similarity to the exemplars of one of the hypothesized categories.

2.03.4 The prototype as a bundle of correlated features

In their paper, Rosch, Mervis, Gray, Johnson, and Boyes-Braem (1976) explicitly described a naturally structured world, consisting of *a priori* categories which were arranged in a hierarchy of increasing abstraction of subordinate, intermediate and superordinate levels. Whilst Putnam's (1977) objects and living entities could belong to only one category at a time, Rosch et al's (1976) could imply a number of categories: *dining-room chair*, a *chair*, *furniture* and *artifactual object* - all simultaneously. Thus, for Putnam, a *chair* might belong to the category *Furniture*, but not be stored in connection with any of the others. In contrast, the concepts on each level within a Roschean hierarchy are made up of correlated feature bundles which are physically similar in appearance, but are distinguished by the characteristic features associated with them. The mental representation of such environmental structures is said to be in terms of prototypical category members.

This 1976 study modified Rosch's previous studies on prototypes in an important way: categories are still ill-defined with fuzzy boundaries, but Rosch now was claiming that there is one level (the basic concept level) at which categories are less fuzzy. Of all the levels in the hierarchy, the structure of the environment is most closely mirrored in what Rosch et al (1976) termed *basic concepts*, because they are said to be more informative, more economical and more useful than at other levels. Such concepts are psychologically "privileged" over others, their level in the conceptual hierarchy allows optimal performance in cognitive activities such as memory, perception, communication. Thus, most subjects' seem to learn and categorize basic concepts more easily, and generate more attributes to a basic concept. This phenomena seems to occur most often with the specific animals and objects found at the intermediate level of the hierarchy. With the increasing generalization above the intermediate level, information is lost; and with the increasing specificity below this level, people's concepts do not seem to acquire many more physical features, nor are they used as often. Hence, Rosch et al. (1976) claimed that optimal subject performance would be found at the intermediate level of the hierarchy, thus defining it as the basic level.

Rosch et al. claimed that basic categories occur because "the perceived world is not an unstructured total set of equiprobable co-occurring attributes. ...the material objects of the world possess high correlational structure" (1976, page 428). She and her colleagues described a very rich ontology, where the environment and its categories might be structured according to common attributes, common function, and common shapes. Their three level hierarchy was based on what they called "the correlational structure of the world", yet their 1976 study results show the effect of variables other than similarity at work, perhaps expertise and personal knowledge, and as a consequence, the basic concept level is not always found at the intermediate level (*chair, table, wardrobe*). Whilst Rosch et al (1976) claimed that optimal subject performance occurred at the intermediate level, thus defining it as the basic concept level, this was not always the case, with subjects performing tasks best at other levels (superordinate or subordinate). Biological concepts were found to be most basic at the superordinate level, not the predicted intermediate level. For example, subjects generated more attributes for *Bird* than for *eagle* or *bald eagle*. On the other hand, the basic level occurred at the (predicted) intermediate level for artifactual categories. The basic level varies between the two types of categories, because subjects seem to know more about concepts like *saw, hammer* than they do about *Tool;* and conversely, know more about *Tree* than they do about *maple, oak.*

One interpretation of this variation in the level of abstraction would be that biological and artifact categories are inherently different, with their respective levels of abstraction being determined by the different types of information required by each category type. This would accord with Rosch et al (1976), who concluded that, in order to be formed, biological categories require more abstraction, and so their superordinate level is the level of most information.

An alternative interpretation is that people might have less experience with such categories, compared with their experience of artifactual categories (Lakoff, 1987)., so that the issue becomes one of the role played by direct experiential knowledge in natural categories. Support for personal knowledge as a factor to take into consideration comes from research involving experts in various fields. The Tzeltal Indians could be said to be experts on the trees found in their jungle. When the researcher's native consultant was asked to name the plants he saw around him, he could name forty or fifty at the intermediate level, including *maple* and *oak.* He also knew more specific names like *sugar maple* and *live oak,* though he did not identify the trees at that level initially. The basic level of classification of *Trees* by Western botanists, as with the Tzeltal Indians, is the intermediate level (Berlin, 1972), compared to the ordinary person's superordinate level (Rosch et al, 1976). In a domain of high expertise, the more information the expert associates with a concept, the more likely it is for that person's basic level to shift from the population norm.

The finding that the basic level can vary according to the amount of knowledge associated with the concept indicates that there is more than universal factors like physical similarity and typicality at work during classification. Berlin (1972, cited in Lakoff, 1987) has suggested that influences such as culture and specialized training also play a part in determining the level of a category which is most distinctive for a person. People from an urban culture that treats *Trees* as the basic level should still have the human capacity to learn to discriminate among trees readily at the intermediate level. Rosch et al's (1976) results, whatever the interpretations, indicate that *perceived* world structure can be constrained by factors other than the environment. In some cases, physical similarity seems to be a minor consideration.

To summarise, all first three variants of the prototype approach propose a simple structural representation for categories, based upon similarity of features, whilst the fourth variant is more complex in that it describes world hierarchical structures of correlated feature bundles. In all four variants, however, physical similarity gives an incomplete account for the coherence of items into one category (rather than another). Komatsu (1992, p.505) gives an excellent example of a Great Dane and a Bedlington terrier who, in appearance, share few similarities although most people would have little difficulty in classifying them as *dogs.* Yet a Bedlington terrier shares as many physical characteristics with a lamb as with a Great Dane, if not more. The family resemblance view of the prototype approach would argue that it is "natural" to partition the world into *dogs* and *lambs,* and that people do this because the summed weights of *terrier* attributes mean that the terrier is more similar to other dogs, than to lambs. Coherence of categories, according to Roschean reasoning, would emerge from the natural world structure as a side-effect (Neisser, 1987). The latest research suggests that similarity may be a byproduct of conceptual coherence, rather than its cause (Komatsu, 1992; Medin, 1989).

2.03.5 Evaluation

The prototype models depend overmuch upon the principle of world structure in formulating their description of membership criteria, so that coherence of items into categories is very loose, since anything can be similar to anything else. In prototype theory, coherence is said to be a side-effect of *perceived* world structure, rather than a direct derivative of it, as a classical approach would claim. The partitioning of objects and creatures in the world into natural categories is said to be constrained by their physical appearance alone. But it has been shown that people extract and retain further information which is unconcerned with physical appearance of the members, such as size of the category, how widely instances vary, and correlations of attributes (Medin, 1989). People will use all this extra knowledge during their categorization decisions, yet it has nothing to do with appearance (Estes, 1986; Flannagan, Fried & Holyoak, 1986; Fried & Holyoak, 1984; Medin, Altom, Edelson & Freko, 1982; Medin & Schaffer, 1978).

Coherence as a property of categories is disregarded by exemplar theories of categorization. In exemplar models, the nature of abstraction is both conservative and tied to the details of specific examples (Medin & Ross, 1989). For example, sensitivity to correlations of properties within a category enables more subtle predictions: from noting that a bird is large, one can predict that it might not sing (Medin, 1989). The exemplar models retain information about category size, instance variability, context, and correlated attributes (Hintzman, 1986; Medin & Schaffer, 1978). A large drawback of these models, however, is that (even more than feature lists or abstract prototype models) specific-exemplar models allow any set of examples to form a category. There is a total lack of constraints on what properties may enter into concepts, or even what constitutes a concept. Consequently, what constitutes *relevance* of information is not even considered. The weakness of exemplar models lies in their loose structure. Because an individual instance is considered to be the prototype, any potential member can match with it on at least a few features.

The abstract prototype model implies that the only information provided by the concept is formal, with no specific details about individual members, such as correlations between features. As a result, if a family resemblance model of categorization were true, inferences would be mistakenly made because too many objects can be categorised as members, simply because they "appear to be" like the prototype. Medin (1989) gives an example about correlated attributes, first looked at by Malt and Smith (1984). Most people have the intuition that small birds are much more likely to sing than large birds. A single summary prototype for *Birds* consisting of the summing of *independent* feature lists, cannot capture this kind of knowledge. Subjects can generate any number of correlated features and this does not support the idea that people reason by using prototypes.

Family resemblance models cannot account for context-effects in categorization and typicality judgments. Roth and Shoben (1983) have shown that context can and does influence judgments of typicality, in that degree-of-typicality of an instance can vary as a function of the context in which the instance is found. In one experiment, they asked subjects how good an example is *tea* of the concept *Beverages.* The subjects responded that it was more typical than *coffee* in the context of secretaries having a break; but less typical when in the context of truck-drivers (who drive at night) taking a break. Perhaps one account which would retain the natural structure position, yet also could account for context-effects would be to propose more specific exemplar concepts: *Truckie Beverages* and *Secretary Beverages.* The exemplar model can do this, but still cannot give a reason for the division into *truckie* and *secretary* in the first place.

To summarise their deficiencies, classical models adhere too strictly to the principle of cognitive economy, so that informativeness about their members is limited, and representations impoverished. Family resemblance models place too much importance upon the principle of world structure of physical similarity. Consequently, the principle of cognitive economy is also inadequately implemented, with information being limited to the similarity of features, and everything else being ignored. Exemplar models allow for too much information but no constraints upon what might be *relevant* information, a necessary function of cognitive economy (Komatsu, 1992). Hierarchical models of correlated feature bundles are less restricted by physical similarity than are family resemblance models, since they include correlational information. However, they also fail to account for people's use of conceptual knowledge during categorization. Finally, the lack of constraints upon choice of what constitutes relevant category knowledge means that classical, family resemblance and exemplar approaches to categorization behaviour are not really giving the full picture of why we have the categories we have; nor why we hold the concepts we do.

2.04 EXPLANATION-BASED MODELS

Rosch's (1983) subsequent solution to the inadequacy of a simple structural representation for categories was based upon the two-stage theory of categorization put forward by Smith, Shoben and Rips (1974), where it was proposed that both characteristic *and* diagnostic features entered into a membership decision. These researchers distinguished between typicality structure and membership criteria in the following way. Concepts were proposed to have *defining* features, which provide a necessary and sufficient determination of set membership; and *characteristic* features which contribute towards the item's degree of typicality. For example *feathers* and *two legs* are among the two defining features used to determine what is a *Bird;* whilst *flying* and *singing* serve to distinguish typical from atypical *Birds.*

According to the two-stage theory, typicality of individual members, and membership rules, are determined in essentially different ways. Although typicality judgments depend upon superficial similarity as described earlier, conceptual relations such as class inclusion, negation, conjunction, and disjunction follow the standard logic of sets, where membership in a set is an all-or-none affair with no uncertainty allowed, and consequently, no gradations in degree of membership.

In 1983, Rosch clarified her theory of category representation. She returned to the ideas expressed in her 1975b paper "Cognitive reference points" and developed them with regard to semantic categories. She came out in favour of dual representation models, which generally represent a division of labour in accounting for different cognitive phenomena. How they do this depends upon the model, some assigning reasoning and identification procedures to the two representations. Rosch's (1983) dual representational model maps onto a distinction between logical and reference point (or analogical) reasoning (Komatsu, 1992).

Briefly, she stated that prototype classification and logical classification need not be mutually exclusive since they are both types of reasoning. The former involves making inferences on the basis of representativeness. A number of events and memories can act as reference points: specific known cases, events or examples contrasted with general knowledge. Categories which do not have determinate boundaries (that is, have ill-defined structure) can only be understood in terms of reference point reasoning, while well-defined categories are subject to both types of reasoning.

The classical and prototype approaches all assume that categorization consists of a similarity comparison between the potential member and the prototype or the conceptual rule In other words, similarity judgments and category judgments are one and the same. Other researchers (Rips, 1989; Murphy & Medin, 1985) have said that similarity-based accounts of people's categorization processes are limited, because people do not always categorize on the basis of an object's appearance alone. Empirical evidence was provided by Rips (1989), who carried out a study where dissociation effects were found between category and similarity judgments.

The fact that the difference between appearance (similarity judgments) and external reality (categorization of objects) is psychologically real has been demonstrated compellingly in a number of transformation studies. The first transformation study was conducted by Carey (1985) who used both adults and children as subjects. The task was to rate a variety of items on their similarity to people. All the item material was animate (for example, worms, flowers, dogs) except for a mechanical monkey that clapped cymbals together when wound up. Adults and children chose the monkey as being most similar to people.

However, neither adults nor children made inductive generalizations of human biological properties to the mechanical monkey. The children (and the adults) showed that they were well aware that the monkey did not have bones, a heart or slept. But they had no trouble attributing these features to mammals, fish and worms which, nevertheless, were rated less similar to humans. Carey (1985) assumed that the monkey had one kind of similarity (a superficial, perceptual kind), whilst the induction task required another kind of similarity (a biologically relevant kind). In short, the mechanical monkey's similarity could not allow it to be categorized as a human, because its similarity was of the wrong kind.

One argument against the theory-based view is that such concepts can be used for categorization only after considerable schooling has been undergone, and factual knowledge learnt (Murphy, 1993b). Because children would not yet possess well-developed theories of the world, their categorizations would be more primitive. Keil (1994) and Carey (1994) have both investigated the constraints and biases which guide name learning. For example, children's beliefs that members of a named category share many properties will support their inferences that novel members also share unobserved attributes.

In 1989, Keil conducted "identity change" studies with children as subjects. The results from the studies eliminated the possibility that physical similarity alone might be the "true" categorization. Pictures of everyday items, both living things and man-made objects, were shown to young children, and their identity given. Then a series of modifications to the object was described, so that the end product looked quite different to the original picture. Keil's (1989) results pointed to some interesting differences in children's perceptions of living things versus man-made objects. Keil (1989) would first present a picture of some animal, for example a raccoon, which would be modified to the extent that it looked and smelled like a skunk. The child's task was to identify the modified object in the last picture, and this involved a choice of criteria between the animal's original identity and its current appearance.

The change from raccoon to skunk was resisted to some degree by all the children of whatever age group (five, seven and nine years), but resistance to change of identity in the man-made objects was not so strong. One example involves the transferral of a coffee-pot into a bird-feeder, where children readily agreed to the change of identity after hearing the physical and functional modifications described and viewing the picture of the end-product. Keil (1989) interpreted his results as support for people's weighting of features to comply with some naive theories of biology, which might involve the belief that an animal's internal structure and genetic history are more central to its identity, than its appearance might be.

Murphy and Medin (1985) have suggested knowledge-based theories of the world as a possible underlying principle. In their paper, a comprehensive Table compares knowledge-based and similarity-based approaches on various aspects of category behaviour. Briefly, people develop their own theories of how the world works, both by learning from others and through their own personal experience. It is not necessary for the item to actually consist of certain features, but rather for the individual to have beliefs about what constitutes the category: the explanatory relationships of its instances to one another and to instances of other categories.

The actual appearance of an object or creature does not determine categorization, but rather, people's beliefs about the genetic structure which underlies that appearance. Putnam's (1975a; 1975b) classical example is that the tiger would still be a tiger, even if it lost its stripes. Kipling (1902), in his *Just So* stories, wrote a story about how the leopard got its spots, which implied that it was already a leopard, even before it gained its spots, which were merely an expression of its "leopardness". Explanation-based theories would argue that the leopard and the tiger have a genetic structure which constitutes and explains their identity, and their appearance is merely an expression of that genetic essence (Medin & Ortony, 1989).

2.05 AIMS AND HYPOTHESES OF THE THESIS

Three properties need to be explained by any theoretical account which aims to be deemed adequate: stability of conceptual representation, coherence of structure among category members, and flexibility of processing during categorization. The quotation below exemplifies some of the properties at work.

> "Do you mean Pendragon's chart of his Pacific Islands?" asked Fanshaw. "*You* thought it was a chart of the Pacific Islands," answered Father Brown. "Put a feather with a fossil and a bit of coral and everyone will think it's a specimen. Put the same feather with a ribbon and an artificial flower and everyone will think it's for a lady's hat. Put the same feather with an ink-bottle, a book and a stack of writing-paper, and most men will swear they've seen a quill pen. So you saw that map among tropic birds and shells and thought it was a map of Pacific Islands. It was the map of this river." (G.K. Chesterton, 1929, page 139-140)

Conceptual stability across people is evident in that "everyone" and "most men" understand what the various groupings will mean (for example, *feather* with *ink-bottle, book* and *writing-paper = quill*). Coherence of the items into a category is also present, if one can create the concept for the category of *Writing Utensils* from the four items given (*ink-bottle, book, writing-paper, quill pen*). Most evident in the above paragraph is the property of flexibility in conceptualization of an object, with *feather* being cross-categorized into a number of different categories. Each context gives salience to different features possessed by *feather,* with the result that it changes identity every time.

2.05.1 Coherence of items into comprehensible categories

The first aim of the thesis is to account for the coherence of items as a comprehensible category unit (rather than some other category). Both classical and prototype models abdicate the responsibility for answering this question, attributing the underlying determinants of categorization to world structure. Human perception of the environment is said to be wholly determined (classical) or merely constrained (prototype) by the partitioning of the world's objects or creatures into natural categories. Category instances are said to naturally belong together because they share most, or all, of the features which characterise the category. However, in limiting constraints upon partitioning of the world to the natural environment, both the classical and prototype view limits the explanatory power of similarity to the appearance-based variety.

In the case of natural categories an appeal to the natural similarity structure of the environment might seem the best explanation for coherence. Yet partitioning of items into a comprehensible category still occurs even though the category-members might be physically dissimilar, as occurs in property types, or ad hoc types. Property category-types have only one attribute defining the category, yet their members still cohere conceptually. Items like *pens, apples* and *cars* are all physically dissimilar except for the one characteristic feature which defines them as belonging to the category of *Red Things.* The fact that such instances can be comprehended as a coherent unit would seem to support the view (held by some classical models) that categories are mere conjunctions of properties selected by the mind and its innate concepts. However, this still does not explain why certain concepts arise rather than others, and it is not guaranteed that useful concepts (for example, useful in predicting events or judging potential members) will evolve, through experience and learning.

2.05.2 Flexibility of relations

The second aim of the thesis is to first consider how certain concepts arise rather than others, and eventually arrive at an explanation for why they arise. In other words, what constitutes a useful concept? For example, why do we see certain feature relations as more relevant or salient than others? What is to prevent a person grouping *cherries* and *meat* together to form the category of *frubidiciousness,* which roughly translates as anything which is *red, juicy* and *edible* (Balzano & McCabe, 1986)? If concepts are to be predictable and stable, achieving a general consensus, the flexibility of their relations with other concepts must be satisfactorily explained.

One example of people's perceived flexibility in similarity relations is provided by Shanon (1988b). He describes a situation where two aunts are viewing their new-born nephew for the first time. Each aunt sees in the baby the facial features resembling one of her forebears, and is quite convinced the new nephew resembles *her* side of the family. In other words, the same face is associated with different features, depending upon which family resemblance prototype it is being compared with. Shanon (1988b) argues that similarity judgments involve constructive processes, with similarity relations between features being determined by the comparison process itself, and by something more than physical similarity.

It is argued here that the most adequate account for concepts and categories is provided by the explanation-based models in the recent research literature. Knowledge-based approaches provide theories/beliefs about how a concept is different from other concepts, thus ensuring that the property of flexibility is incorporated into the account. They also provide theories/beliefs which explain the causal relations amongst members of a category, thus including the property of coherence of members into a category. In short, a theory-based concept can deal with inter-conceptual and intra-conceptual relations, as described below:

> The explanation-, or knowledge-, based view of concepts tries to explain the simultaneous properties of coherence and flexibility by arguing that the specification of a concept includes information about how that concept is related to other concepts (or how its instances relate to other objects) and about the relationships - especially the functional, causal, or explanatory relationships - that hold among the attributes associated with its instances. For example, the concept piano may include the information that people typically sit on a bench to play it (i.e., a relationship to other concepts). The concept bird may include the information that birds have a certain genetic structure that under normal conditions expresses itself in having wings, feathers, and so on (i.e. relationships among attributes associated with its instances). (Komatsu, 1992, p. 515).

Knowledge-based approaches can deal with such relations because the conceptual core does not contain merely formal rules (as in classical models) or featural structures (as in prototype models), but also contains content of meaningful information (Murphy, 1993b). It is this meaningful information which both coheres certain items as members of a category, and which provides constraints (by specifying *relevant* information) upon conceptual relations with other concepts. By providing knowledge-based constraints upon relations with other concepts, this view can account for the *flexibility* of concepts.

Neither classical nor prototype theories can specify what constitutes relevant criteria, and so they are not really saying anything about why certain concepts arise rather than others. The importance of relevant knowledge, as a constraint upon both relational flexibility and item-coherence, has been much emphasised by Murphy and Medin (1985). Otherwise, anything might be similar to anything else, and without the knowledge of what constitutes relevance for a concept and its category of items, flexibility might degenerate into unconstrained instability of concepts (Wattenmaker, Nakamura & Medin, 1988).

Another factor in favour of theory-based models of categorization is that they give a more active account of the person's cognitive functioning, rather than processes which are totally passive or driven by perception of appearance alone. Allowance is made for human interpretation of the environment. The focus on the structure of the environment which began with Rosch's studies in the early seventies came at the cost of neglecting the nature of the person who forms and uses categories (Gardner, 1987).

Because the classical and prototype models do not fulfil the principles of world structure and cognitive economy, they give weak accounts of conceptual representation, structure and processing of categories. The classical approach falls short in its computational *processes*, which are too rigid, because they assume one-to-one (or isomorphic) relations between an item and its category. The prototype approach is inadequate where *structure* is concerned. The family resemblance models have an impoverished *representation*, in that they are constituted solely of structural representation of feature lists or feature relations, which does not include content. The exemplar prototype models are also impoverished because they lack constraints, in other words their *structure* is too loose, and anything can be similar to anything else.

2.05.3 Hypotheses

Two questions to be addressed by the thesis were posed at the beginning of chapter 1. The first was concerned with how certain concepts arise, and from whence their stability might be derived. The second asked why we have the categories we do. The thesis will look at a number of possible answers to the questions, including the classical and the similarity-based approaches. It is hypothesised that we have the categories we do (and not others) because of the concepts we construct. These concepts are constructed for the purpose of explaining the world around us, using the theories/beliefs which constitute our background knowledge about the world.

It is suggested here, and will be empirically tested, that classical and similarity-based representations processes and structure are inadequate as accounts of concept and category behaviour. Also, it is argued that explanation-based models provide the best account available of concept and category behaviour, because they provide the informational constraints necessary for the properties of flexibility and coherence to be incorporated into the theory.

More specifically, it is hypothesised that (a) the stability of concepts is not due to their representation in a formal classical mode; (b) the representation, structure and process of categories as described by the classical and prototype models is inadequate; and (c) the most relevant information contained in concepts will be concerned with personal details, such as goals, needs, purposes.

2.06 EMPIRICAL STUDIES OF THE THESIS

Experiment 1 is concerned with whether people represent a category's members at all. The more recent exponents of the classical approach to categories claim that people have simple, innate "operators" which directly and passively interpret the structure of the outside world (Anderson, 1991a). They argue that computational algorithms are sufficient representations. Empirically, classical and prototype models will be compared as to whether members in three category-types (superordinate, property and ad hoc types) are represented in a statistical fashion, or whether a normative representation (shared by participants in the experiment) is involved.

Experiment 2 is concerned with whether people use the same representations for different category-types. It compares category processes, representation and structure for the three category-types used in experiment 1, using both the normative stimuli from the previous experiment and idiosyncratic stimuli generated by individual subjects. One question concerns the role of physical similarity in some category-types whose members are not physically alike, such as ad hoc categories. Surface similarity cannot be said to play an important role, if any, in ad hoc category-types (Neisser, 1987). Yet they still show certain features which the prototype approach claims are peculiar to natural categories. These include coherence as a psychological unit, internal membership gradience and typicality effects (Barsalou, 1983; Barsalou & Ross, 1986; Hampton, 1981). Other category-types which remain unexplained are those whose member items all share the one and only feature (be it functional, perceptual, or conceptual), as is the case for property category-types.

Whilst Rosch might have changed her mind about prototypes as representations of categories, her research has proved to be a source of interest for later researchers. See Smith and Medin (1981) for a survey of past research which was based upon these assumptions. Many experiments have been based upon a number of similarity-based assumptions which can be summed up as saying that prototype effects reflect something *direct* about the nature of human categorization and representation. Experiment 2 tests these assumptions of the similarity-based approach by investigating how well the representation, structure and process of the three category-types match such assumptions.

Experiment 3 examines whether people hold more than one representation for the same concept. If similarity is not a sufficient basis for categorization and diagnostic representation, what is missing? The study investigates what it is that influences the *perceived* similarity in a feature, that is, what is relevant knowledge where categorization judgments are concerned. It is possible that we have separate representations of the same concept: one representation for the individual members' similarity of appearance and a second representation for the members' essential reality, the conceptual core criterion. Philosophically, the issue is concerned with whether physical appearance and essential reality are one and the same; or whether we use our beliefs about reality to interpret physical appearance.

A related question concerns relevant criteria. It is possible that people use more than physical similarity as a basis for categorization, in which case what features or naive theories do they consider to be most relevant? For example, would functional features be more relevant than physically similar features when dealing with artifactual categories? Experiment 3 will look at how much influence (if any) the various criteria have on people's categorization, similarity and typicality judgments of artifactual versus biological concepts.