David McNeill, at the University of Chicago
Hilary Putnam, at Harvard University
Naomi Quinn, at Duke University
John Robert Ross, at the Massachusetts Institute of Technology
David Zubin, at the State University of New York at Buffalo

I would also like to thank R. M. W. Dixon and Annette Schmidt of the Australian National University for providing me with a lengthy discussion of their research on Dyirbal categorization, as well as Pamela Downing and Haruo Aoki, who provided me with details about Japanese classifiers. Mark Johnson and Hilary Putnam have been extremely helpful in discussing philosophical issues, especially their recent work. The philosophical views put forth here have been worked out in collaboration with Johnson over many years. Robert Solovay and Saunders Mac Lane provided enormously useful discussions of the foundations of mathematics. Extensive comments on drafts of the manuscript have been provided by Jay Atlas, Lawrence Barsalou, Claudia Brugman, Michele Emanatian, Charles Fillmore, Jim Greeno, Mark Johnson, Paul Kay, Zoltán Kövecses, Robert McCauley, James D. McCawley, Carolyn Mervis, Ulric Neisser, Eleanor Rosch, Edward Smith, Robert Wilensky. Sustenance of extraordinary quality was provided by Cafe Fanny in Berkeley.

This research would not have been possible without grants from the National Science Foundation (grant no. BNS-8310445), the Sloan Foundation, and the Committee on Research of the University of California at Berkeley. I would especially like to thank Paul Chapin at NSF and Eric Wanner at Sloan.

Large projects like this cannot be completed without enormous sacrifices on the home front. Claudia Brugman and Andy Lakoff have put up with my unavailability for longer than I would like to think. I thank them for their patience and perseverance with all my heart.

Berkeley, California
July, 1985

Preface

Cognitive science is a new field that brings together what is known about the mind from many academic disciplines: psychology, linguistics, anthropology, philosophy, and computer science. It seeks detailed answers to such questions as: What is reason? How do we make sense of our experience? What is a conceptual system and how is it organized? Do all people use the same conceptual system? If so, what is that system? If not, exactly what is there that is common to the way all human beings think? The questions aren't new, but some recent answers are.

This book is about the traditional answers to these questions and about recent research that suggests new answers. On the traditional view, reason is abstract and disembodied. On the new view, reason has a bodily basis. The traditional view sees reason as literal, as primarily about propositions that can be objectively either true or false. The new view takes imaginative aspects of reason—metaphor, metonymy, and mental imagery—as central to reason, rather than as a peripheral and inconsequential adjunct to the literal.

The traditional account claims that the capacity for meaningful thought and for reason is abstract and not necessarily embodied in any organism. Thus, meaningful concepts and rationality are *transcendental*, in the sense that they transcend, or go beyond, the physical limitations of any organism. Meaningful concepts and abstract reason may happen to be embodied in human beings, or in machines, or in other organisms—but they exist abstractly, independent of any particular embodiment. In the new view, meaning is a matter of what is meaningful to thinking, functioning beings. The nature of the thinking organism and the way it functions in its environment are of central concern to the study of reason.

Both views take categorization as the main way that we make sense of experience. Categories on the traditional view are characterized solely by the properties shared by their members. That is, they are characterized

(*a*) independently of the bodily nature of the beings doing the categorizing and (*b*) literally, with no imaginative mechanisms (metaphor, metonymy, and imagery) entering into the nature of categories. In the new view, our bodily experience and the way we use imaginative mechanisms are central to how we construct categories to make sense of experience.

Cognitive science is now in transition. The traditional view is hanging on, although the new view is beginning to take hold. Categorization is a central issue. The traditional view is tied to the classical theory that categories are defined in terms of common properties of their members. But a wealth of new data on categorization appears to contradict the traditional view of categories. In its place there is a new view of categories, what Eleanor Rosch has termed *the theory of prototypes and basic-level categories*. We will be surveying that data and its implications.

The traditional view is a philosophical one. It has come out of two thousand years of philosophizing about the nature of reason. It is still widely believed despite overwhelming empirical evidence against it. There are two reasons. The first is simply that it is traditional. The accumulated weight of two thousand years of philosophy does not go away overnight. We have all been educated to think in those terms. The second reason is that there has been, until recently, nothing approaching a well-worked-out alternative that preserves what was correct in the traditional view while modifying it to account for newly discovered data. This book will also be concerned with describing such an alternative.

We will be calling the traditional view *objectivism* for the following reason: Modern attempts to make it work assume that rational thought consists of the manipulation of abstract symbols and that these symbols get their meaning via a correspondence with the world, *objectively construed*, that is, independent of the understanding of any organism. A collection of symbols placed in correspondence with an objectively structured world is viewed as a *representation* of reality. On the objectivist view, *all* rational thought involves the manipulation of abstract symbols which are given meaning only via conventional correspondences with things in the external world.

Among the more specific objectivist views are the following:

- Thought is the mechanical manipulation of abstract symbols.
- The mind is an abstract machine, manipulating symbols essentially in the way a computer does, that is, by algorithmic computation.
- Symbols (e.g., words and mental representations) get their meaning via correspondences to things in the external world. All meaning is of this character.

- Symbols that correspond to the external world are *internal representations of external reality*.
- Abstract symbols may stand in correspondence to things in the world independent of the peculiar properties of any organisms.
- Since the human mind makes use of internal representations of external reality, the mind is *a mirror of nature*, and correct reason mirrors the logic of the external world.
- It is thus incidental to the nature of meaningful concepts and reason that human beings have the bodies they have and function in their environment in the way they do. Human bodies may play a role in *choosing* which concepts and which modes of transcendental reason human beings actually employ, but they play no essential role in *characterizing* what constitutes a concept and what constitutes reason.
- Thought is *abstract* and *disembodied*, since it is independent of any limitations of the human body, the human perceptual system, and the human nervous system.
- Machines that do no more than mechanically manipulate symbols that correspond to things in the world are capable of meaningful thought and reason.
- Thought is *atomistic*, in that it can be completely broken down into simple "building blocks"—the symbols used in thought—which are combined into complexes and manipulated by rule.
- Thought is *logical* in the narrow technical sense used by philosophical logicians; that is, it can be modeled accurately by systems of the sort used in mathematical logic. These are abstract symbol systems defined by general principles of symbol manipulation and mechanisms for interpreting such symbols in terms of "models of the world."

Though such views are by no means shared by all cognitive scientists, they are nevertheless widespread, and in fact so common that many of them are often assumed to be true without question or comment. Many, perhaps even most, contemporary discussions of the mind as a computing machine take such views for granted.

The idea of a *category* is central to such views. The reason is that most symbols (i.e., words and mental representations) do not designate particular things or individuals in the world (e.g., Rickey Henderson or the Golden Gate Bridge). Most of our words and concepts designate categories. Some of these are categories of things or beings in the physical world—chairs and zebras, for example. Others are categories of activities and abstract things—singing and songs, voting and governments, etc. To a very large extent, the objectivist view of language and thought rests on

the nature of categories. On the objectivist view, things are in the same category if and only if they have certain properties in common. Those properties are necessary and sufficient conditions for defining the category.

On the objectivist view of meaning, the symbols used in thought get their meaning via their correspondence with things—particular things or categories of things—in the world. Since categories, rather than individuals, matter most in thought and reason, a category must be the sort of thing that can fit the objectivist view of mind in general. All conceptual categories must be symbols (or symbolic structures) that can designate categories in the real world, or in some possible world. And the world must come divided up into categories of the right kind so that symbols and symbolic structures can refer to them. "Categories of the right kind" are classical categories, categories defined by the properties common to all their members.

In recent years, conceptual categories have been studied intensively and in great detail in a number of the cognitive sciences—especially anthropology, linguistics, and psychology. The evidence that has accumulated is in conflict with the objectivist view of mind. Conceptual categories are, on the whole, very different from what the objectivist view requires of them. That evidence suggests a very different view, not only of categories, but of human reason in general:

- Thought is *embodied*, that is, the structures used to put together our conceptual systems grow out of bodily experience and make sense in terms of it; moreover, the core of our conceptual systems is directly grounded in perception, body movement, and experience of a physical and social character.
- Thought is *imaginative*, in that those concepts which are not directly grounded in experience employ metaphor, metonymy, and mental imagery—all of which go beyond the literal mirroring, or *representation*, of external reality. It is this imaginative capacity that allows for "abstract" thought and takes the mind beyond what we can see and feel. The imaginative capacity is also embodied—indirectly—since the metaphors, metonymies, and images are based on experience, often bodily experience. Thought is also imaginative in a less obvious way: every time we categorize something in a way that does not mirror nature, we are using general human imaginative capacities.
- Thought has *gestalt properties* and is thus not atomistic; concepts have an overall structure that goes beyond merely putting together conceptual "building blocks" by general rules.
- Thought has an *ecological structure*. The efficiency of cognitive processing, as in learning and memory, depends on the overall structure of the conceptual system and on what the concepts mean. Thought is thus more than just the mechanical manipulation of abstract symbols.
- Conceptual structure can be described using *cognitive models* that have the above properties.
- The theory of cognitive models incorporates what was right about the traditional view of categorization, meaning, and reason, while accounting for the empirical data on categorization and fitting the new view overall.

I will refer to the new view as *experiential realism* or alternatively as *experientialism*. The term *experiential realism* emphasizes what experientialism shares with objectivism: (*a*) a commitment to the existence of the real world, (*b*) a recognition that reality places constraints on concepts, (*c*) a conception of truth that goes beyond mere internal coherence, and (*d*) a commitment to the existence of stable knowledge of the world.

Both names reflect the idea that thought fundamentally grows out of embodiment. "Experience" here is taken in a broad rather than a narrow sense. It includes everything that goes to make up actual or potential experiences of either individual organisms or communities of organisms —not merely perception, motor movement, etc., but *especially* the internal genetically acquired makeup of the organism and the nature of its interactions in both its physical and its social environments.

Experientialism is thus defined in contrast with objectivism, which holds that the characteristics of the organism have nothing essential to do with concepts or with the nature of reason. On the objectivist view, human reason is just a limited form of transcendental reason. The only roles accorded to the body are (*a*) to provide access to abstract concepts, (*b*) to provide "wetware," that is, a biological means of mimicking patterns of transcendental reason, and (*c*) to place limitations on possible concepts and forms of reason. On the experientialist view, reason is made possible by the body—that includes abstract and creative reason, as well as reasoning about concrete things. Human reason is not an instantiation of transcendental reason; it grows out of the nature of the organism and all that contributes to its individual and collective experience: its genetic inheritance, the nature of the environment it lives in, the way it functions in that environment, the nature of its social functioning, and the like.

The issue is this:

> Do meaningful thought and reason concern merely the manipulation of abstract symbols and their correspondence to an objective reality, independent of any embodiment (except, perhaps, for limitations imposed by the organism)?

Or do meaningful thought and reason essentially concern the nature of the organism doing the thinking—including the nature of its body, its interactions in its environment, its social character, and so on?

Though these are highly abstract questions, there does exist a body of evidence that suggests that the answer to the first question is no and the answer to the second is yes. That is a significant part of what this book is about.

Why does all this matter? It matters for our understanding of who we are as human beings and for all that follows from that understanding. The capacity to reason is usually taken as defining what human beings are and as distinguishing us from other things that are alive. If we understand reason as being disembodied, then our bodies are only incidental to what we are. If we understand reason as mechanical—the sort of thing a computer can do—then we will devalue human intelligence as computers get more efficient. If we understand rationality as the capacity to mirror the world external to human beings, then we will devalue those aspects of the mind that can do infinitely more than that. If we understand reason as merely literal, we will devalue art.

How we understand the mind matters in all these ways and more. It matters for what we value in ourselves and others—for education, for research, for the way we set up human institutions, and most important for what counts as a humane way to live and act. If we understand reason as embodied, then we will want to understand the relationship between the mind and the body and to find out how to cultivate the embodied aspects of reason. If we fully appreciate the role of the imaginative aspects of reason, we will give them full value, investigate them more thoroughly, and provide better education in using them. Our ideas about what people can learn and should be learning, as well as what they should be doing with what they learn, depend on our concept of learning itself. It is important that we have discovered that learning for the most part is neither rote learning nor the learning of mechanical procedures. It is important that we have discovered that rational thought goes well beyond the literal and the mechanical. It is important because our ideas about how human minds should be employed depend on our ideas of what a human mind is.

It also matters in a narrower but no less important way. Our understanding of what reason is guides our current research on the nature of reason. At present, that research is expanding faster than at any time in history. The research choices made now by the community of cognitive scientists will shape our view of mind for a long time to come. We are at present at an important turning point in the history of the study of the mind. It is vital that the mistaken views about the mind that have been with us for two thousand years be corrected.

This book attempts to bring together some of the evidence for the view that reason is embodied and imaginative—in particular, the evidence that comes from the study of the way people categorize. Conceptual systems are organized in terms of categories, and most if not all of our thought involves those categories. The objectivist view rests on a theory of categories that goes back to the ancient Greeks and that even today is taken for granted as being not merely true, but obviously and unquestionably true. Yet contemporary studies of the way human beings actually categorize things suggest that categorization is a rather different and more complex matter.

What is most interesting to me about these studies is that they seem to provide evidence for the experientialist view of human reason and against the objectivist view. Taken one by one, such studies are things only scholars could care about, but taken as a whole, they have something magnificent about them: evidence that the mind is more than a mere mirror of nature or a processor of symbols, that it is not incidental to the mind that we have bodies, and that the capacity for understanding and meaningful thought goes beyond what any machine can do.

CHAPTER 1

The Importance of Categorization

Many readers, I suspect, will take the title of this book as suggesting that women, fire, and dangerous things have something in common—say, that women are fiery and dangerous. Most feminists I've mentioned it to have loved the title for that reason, though some have hated it for the same reason. But the chain of inference—from conjunction to categorization to commonality—is the norm. The inference is based on the common idea of what it means to be in the same category: things are categorized together on the basis of what they have in common. The idea that categories are defined by common properties is not only our everyday folk theory of what a category is, it is also the principal technical theory—one that has been with us for more than two thousand years.

The classical view that categories are based on shared properties is not entirely wrong. We often do categorize things on that basis. But that is only a small part of the story. In recent years it has become clear that categorization is far more complex than that. A new theory of categorization, called *prototype theory*, has emerged. It shows that human categorization is based on principles that extend far beyond those envisioned in the classical theory. One of our goals is to survey the complexities of the way people really categorize. For example, the title of this book was inspired by the Australian aboriginal language Dyirbal, which has a category, *balan*, that actually includes women, fire, and dangerous things. It also includes birds that are *not* dangerous, as well as exceptional animals, such as the platypus, bandicoot, and echidna. This is not simply a matter of categorization by common properties, as we shall see when we discuss Dyirbal classification in detail.

Categorization is not a matter to be taken lightly. There is nothing more basic than categorization to our thought, perception, action, and speech. Every time we see something as a *kind* of thing, for example, a tree, we are categorizing. Whenever we reason about *kinds* of things—chairs, nations, illnesses, emotions, any kind of thing at all—we

are employing categories. Whenever we intentionally perform any *kind* of action, say something as mundane as writing with a pencil, hammering with a hammer, or ironing clothes, we are using categories. The particular action we perform on that occasion is a *kind* of motor activity (e.g., writing, hammering, ironing), that is, it is in a particular category of motor actions. They are never done in exactly the same way, yet despite the differences in particular movements, they are all movements of a kind, and we know how to make movements of that kind. And any time we either produce or understand any utterance of any reasonable length, we are employing dozens if not hundreds of categories: categories of speech sounds, of words, of phrases and clauses, as well as conceptual categories. Without the ability to categorize, we could not function at all, either in the physical world or in our social and intellectual lives. An understanding of how we categorize is central to any understanding of how we think and how we function, and therefore central to an understanding of what makes us human.

Most categorization is automatic and unconscious, and if we become aware of it at all, it is only in problematic cases. In moving about the world, we automatically categorize people, animals, and physical objects, both natural and man-made. This sometimes leads to the impression that we just categorize things as they are, that things come in natural kinds, and that our categories of mind naturally fit the kinds of things there are in the world. But a large proportion of our categories are not categories of *things*; they are categories of abstract entities. We categorize events, actions, emotions, spatial relationships, social relationships, and abstract entities of an enormous range: governments, illnesses, and entities in both scientific and folk theories, like electrons and colds. Any adequate account of human thought must provide an accurate theory for *all* our categories, both concrete and abstract.

From the time of Aristotle to the later work of Wittgenstein, categories were thought be well understood and unproblematic. They were assumed to be abstract containers, with things either inside or outside the category. Things were assumed to be in the same category if and only if they had certain properties in common. And the properties they had in common were taken as defining the category.

This classical theory was not the result of empirical study. It was not even a subject of major debate. It was a philosophical position arrived at on the basis of a priori speculation. Over the centuries it simply became part of the background assumptions taken for granted in most scholarly disciplines. In fact, until very recently, the classical theory of categories was not even thought of as a *theory*. It was taught in most disciplines not as an empirical hypothesis but as an unquestionable, definitional truth.

In a remarkably short time, all that has changed. Categorization has moved from the background to center stage because of empirical studies in a wide range of disciplines. Within cognitive psychology, categorization has become a major field of study, thanks primarily to the pioneering work of Eleanor Rosch, who made categorization an issue. She focused on two implications of the classical theory:

> First, if categories are defined only by properties that all members share, then no members should be better examples of the category than any other members.
>
> Second, if categories are defined only by properties inherent in the members, then categories should be independent of the peculiarities of any beings doing the categorizing; that is, they should not involve such matters as human neurophysiology, human body movement, and specific human capacities to perceive, to form mental images, to learn and remember, to organize the things learned, and to communicate efficiently.

Rosch observed that studies by herself and others demonstrated that categories, in general, have best examples (called "prototypes") and that all of the specifically human capacities just mentioned do play a role in categorization.

In retrospect, such results should not have been all that surprising. Yet the specific details sent shock waves throughout the cognitive sciences, and many of the reverberations are still to be felt. Prototype theory, as it is evolving, is changing our idea of the most fundamental of human capacities—the capacity to categorize—and with it, our idea of what the human mind and human reason are like. Reason, in the West, has long been assumed to be disembodied and abstract—distinct on the one hand from perception and the body and culture, and on the other hand from the mechanisms of imagination, for example, metaphor and mental imagery.

In this century, reason has been understood by many philosophers, psychologists, and others as roughly fitting the model of formal deductive logic:

> Reason is the mechanical manipulation of abstract symbols which are meaningless in themselves, but can be given meaning by virtue of their capacity to refer to things either in the actual world or in possible states of the world.

Since the digital computer works by symbol manipulation and since its symbols can be interpreted in terms of a data base, which is often viewed as a partial model of reality, the computer has been taken by many as essentially possessing the capacity to reason. This is the basis of the contem-

porary mind-as-computer metaphor, which has spread from computer science and cognitive psychology to the culture at large.

Since we reason not just about individual things or people but about categories of things and people, categorization is crucial to every view of reason. Every view of reason must have an associated account of categorization. The view of reason as the *disembodied* manipulation of abstract symbols comes with an implicit theory of categorization. It is a version of the classical theory in which categories are represented by sets, which are in turn defined by the properties shared by their members.

There is a good reason why the view of reason as disembodied symbol-manipulation makes use of the classical theory of categories. If symbols in general can get their meaning only through their capacity to correspond to things, then *category* symbols can get their meaning only through a capacity to correspond to *categories* in the world (the real world or some possible world). Since the symbol-to-object correspondence that defines meaning in general must be independent of the peculiarities of the human mind and body, it follows that the symbol-to-category correspondence that defines meaning for category symbols must also be independent of the peculiarities of the human mind and body. To accomplish this, categories must be seen as existing in the world independent of people and defined only by the characteristics of their members and not in terms of any characteristics of the human. The classical theory is just what is needed, since it defines categories only in terms of shared properties of the *members* and not in terms of the peculiarities of human understanding.

To question the classical view of categories in a fundamental way is thus to question the view of reason as disembodied symbol-manipulation and correspondingly to question the most popular version of the mind-as-computer metaphor. Contemporary prototype theory does just that—through detailed empirical research in anthropology, linguistics, and psychology.

The approach to prototype theory that we will be presenting here suggests that human categorization is essentially a matter of both human experience and imagination—of perception, motor activity, and culture on the one hand, and of metaphor, metonymy, and mental imagery on the other. As a consequence, human reason crucially depends on the same factors, and therefore cannot be characterized merely in terms of the manipulation of abstract symbols. Of course, certain aspects of human reason can be isolated artificially and modeled by abstract symbol-manipulation, just as some part of human categorization does fit the classical theory. But we are interested not merely in some artificially isolatable subpart of the human capacity to categorize and reason, but in the full range of that capacity. As we shall see, those aspects of categorization that do fit the classical theory are special cases of a general theory of cognitive models, one that permits us to characterize the experiential and imaginative aspects of reason as well.

To change the very concept of a category is to change not only our concept of the mind, but also our understanding of the world. Categories are categories *of* things. Since we understand the world not only in terms of individual things but also in terms of *categories* of things, we tend to attribute a real existence to those categories. We have categories for biological species, physical substances, artifacts, colors, kinsmen, and emotions and even categories of sentences, words, and meanings. We have categories for everything we can think about. To change the concept of *category* itself is to change our understanding of the world. At stake is our understanding of everything from what a biological species is (see chap. 12) to what a word is (see case study 2).

The evidence we will be considering suggests a shift from classical categories to prototype-based categories defined by cognitive models. It is a change that implies other changes: changes in the concepts of truth, knowledge, meaning, rationality—even grammar. A number of familiar ideas will fall by the wayside. Here are some that will have to be left behind:

- Meaning is based on truth and reference; it concerns the relationship between symbols and things in the world.
- Biological species are natural kinds, defined by common essential properties.
- The mind is separate from, and independent of, the body.
- Emotion has no conceptual content.
- Grammar is a matter of pure form.
- Reason is transcendental, in that it transcends—goes beyond—the way human beings, or any other kinds of beings, happen to think. It concerns the inferential relationships among all possible concepts in this universe or any other. Mathematics is a form of transcendental reason.
- There is a correct, God's eye view of the world—a single correct way of understanding what is and is not true.
- All people think using the same conceptual system.

These ideas have been part of the superstructure of Western intellectual life for two thousand years. They are tied, in one way or another, to the classical concept of a category. When that concept is left behind, the others will be too. They need to be replaced by ideas that are not only more accurate, but more humane.

Many of the ideas we will be arguing against, on empirical grounds, have been taken as part of what *defines* science. One consequence of this study will be that certain common views of science will seem too narrow. Consider, for example, scientific rigor. There is a narrow view of science that considers as rigorous only hypotheses framed in first-order predicate calculus with a standard model-theoretic interpretation, or some equivalent system, say a computer program using primitives that are taken as corresponding to an external reality. Let us call this the predicate calculus (or "PC") view of scientific theorizing. The PC view characterizes explanations only in terms of deductions from hypotheses, or correspondingly, in terms of computations. Such a methodology not only claims to be rigorous in itself, it also claims that no other approach can be sufficiently precise to be called scientific. The PC view is prevalent in certain communities of linguists and cognitive psychologists and enters into many investigations in the cognitive sciences.

Such a view of science has long been discredited among philosophers of science (for example, see Hanson 1961, Hesse 1963, Kuhn 1970, 1977, and Feyerabend 1975). As we will see (chaps. 11–20), the PC view is especially inappropriate in the cognitive sciences since it *assumes* an a priori view of categorization, namely, the classical theory that categories are sets defined by common properties of objects. Such an assumption makes it impossible to ask, as an empirical question, whether the classical view of categorization is correct. The classical view is assumed to be correct, because it is built into classical logic, and hence into the PC view. Thus, we sometimes find circular arguments about the nature of categorization that are of the following form:

Premise (often hidden): The PC view of scientific rigor is correct.

. . .

. . .

. . .

Conclusion: Categories are classical.

The conclusion is, of course, presupposed by the premise. To avoid vacuity, the empirical study of categorization cannot take the PC view of scientific rigor for granted.

A central goal of cognitive science is to discover what reason is like and, correspondingly, what categories are like. It is therefore especially important for the study of cognitive science not to assume the PC view, which presupposes an a priori answer to such empirical questions. This, of course, does not mean that one cannot be rigorous or precise. It only means that rigor and precision must be characterized in another way—a

way that does not stifle the empirical study of the mind. We will suggest such a way in chapter 17.

The PC view of rigor leads to rigor mortis in the study of categorization. It leads to a view of the sort proposed by Osherson and Smith (1981) and Armstrong, Gleitman, and Gleitman (1983) and discussed in chapter 9 below, namely, that the classical view of categorization is correct and the enormous number of phenomena that do not accord with it are either due to an "identification" mechanism that has nothing to do with reason or are minor "recalcitrant" phenomena. As we go through this book, we will see that there seem to be more so-called recalcitrant phenomena than there are phenomena that work by the classical view.

This book surveys a wide variety of rigorous empirical studies of the nature of human categorization. In concluding that categorization is not classical, the book implicitly suggests that the PC view of scientific rigor is itself not scientifically valid. The result is not chaos, but an expanded perspective on human reason, one which by no means requires imprecision or vagueness in scientific inquiry. The studies cited, for example, those by Berlin, Kay, Ekman, Rosch, Tversky, Dixon, and many others, more than meet the prevailing standards of scientific rigor and accuracy, while challenging the conception of categories presupposed by the PC view of rigor. In addition, the case studies presented below in Book II are intended as examples of empirical research that meet or exceed the prevailing standards. In correcting the classical view of categorization, such studies serve to raise the general standards of scientific accuracy in the cognitive sciences.

The view of categorization that I will be presenting has not arisen all at once. It has developed through a number of intermediate stages that lead up to the cognitive model approach. An account of those intermediate steps begins with the later philosophy of Ludwig Wittgenstein and goes up through the psychological research of Eleanor Rosch and her associates.

CHAPTER 4

Idealized Cognitive Models

Sources of Prototype Effects

The main thesis of this book is that we organize our knowledge by means of structures called *idealized cognitive models*, or ICMs, and that category structures and prototype effects are by-products of that organization. The ideas about cognitive models that we will be making use of have developed within cognitive linguistics and come from four sources: Fillmore's frame semantics (Fillmore 1982*b*), Lakoff and Johnson's theory of metaphor and metonymy (Lakoff and Johnson 1980), Langacker's cognitive grammar (Langacker 1986), and Fauconnier's theory of mental spaces (Fauconnier 1985). Fillmore's frame semantics is similar in many ways to schema theory (Rumelhart 1975), scripts (Schank and Abelson 1977), and frames with defaults (Minsky 1975). Each ICM is a complex structured whole, a gestalt, which uses four kinds of structuring principles:

- propositional structure, as in Fillmore's frames
- image-schematic structure, as in Langacker's cognitive grammar
- metaphoric mappings, as described by Lakoff and Johnson
- metonymic mappings, as described by Lakoff and Johnson

Each ICM, as used, structures a mental space, as described by Fauconnier.

Probably the best way to provide an idea of what ICMs are and how they work in categorization is to go through examples. Let us begin with Fillmore's concept of a *frame*. Take the English word *Tuesday*. *Tuesday* can be defined only relative to an idealized model that includes the natural cycle defined by the movement of the sun, the standard means of characterizing the end of one day and the beginning of the next, and a larger seven-day calendric cycle—the week. In the idealized model, the week is a whole with seven parts organized in a linear sequence; each part is called a *day*, and the third is *Tuesday*. Similarly, the concept *weekend* requires a notion of a *work week* of five days followed by a break of two days, superimposed on the seven-day calendar.

Our model of a week is idealized. Seven-day weeks do not exist objectively in nature. They are created by human beings. In fact, not all cultures have the same kinds of weeks. Consider, for example, the Balinese calendric system:

> The two calendars which the Balinese employ are a lunar-solar one and one built around the interaction of independent cycles of day-names, which I shall call "permutational." The permutational calendar is by far the most important. It consists of ten different cycles of day-names, following one another in a fixed order, after which the first day-name appears and the cycle starts over. Similarly, there are nine, eight, seven, six, five, four, three, two, and even—the ultimate of a "contemporized" view of time—one day-name cycles. The names in each cycle are also different, and the cycles run concurrently. That is to say, any given day has, at least in theory, ten different names simultaneously applied to it, one from each of the ten cycles. Of the ten cycles, only those containing five, six, and seven day-names are of major cultural significance. . . . The outcome of all this wheels-within-wheels computation is a view of time as consisting of ordered sets of thirty, thirty-five, forty-two and two hundred and ten quantum units ("days"). . . . To identify a day in the forty-two-day set—and thus assess its practical and/or religious significance—one needs to determine its place, that is, its name in the six-name cycle (say *Ariang*) and in the seven-day cycle (say *Boda*): the day is Boda-Ariang, and one shapes one's actions accordingly. To identify a day in the thirty-five day set, one needs its place and name in the five-name cycle (for example, *Klion*) and in the seven-: for example, *Boda-Klion*. . . . For the two-hundred-and-ten-day set, unique determination demands names from all three weeks: for example, *Boda-Ariang-Klion*, which, it so happens, is the day on which the most important Balinese holiday, Galungan, is celebrated. (Geertz 1973, pp. 392–93)

Thus, a characterization of *Galungan* in Balinese requires a complex ICM which superimposes three week-structures—one five-day, one six-day, and one seven-day. In the cultures of the world, such idealized cognitive models can be quite complex.

The Simplest Prototype Effects

In general, any element of a cognitive model can correspond to a conceptual category. To be more specific, suppose schema theory in the sense of Rumelhart (1975) were taken as characterizing propositional models. Each schema is a network of nodes and links. Every node in a schema would then correspond to a conceptual category. The properties of the category would depend on many factors: the role of that node in the given

schema, its relationship to other nodes in the schema, the relationship of that schema to other schemas, and the overall interaction of that schema with other aspects of the conceptual system. As we will see, there is more to ICMs than can be represented in schema theory. But at least those complexities do arise. What is particularly interesting is that even if one set up schema theory as one's theory of ICMs, and even if the categories defined in those schemas were classical categories, there would still be prototype effects—effects that would arise from the interaction of the given schema with other schemas in the system.

A clear example of this has been given by Fillmore (1982*a*). The example is a classic: the category defined by the English word *bachelor.*

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

In other words, *bachelor* is defined with respect to an ICM in which there is a human society with (typically monogamous) marriage, and a typical marriageable age. The idealized model says nothing about the existence of priests, "long-term unmarried couplings," homosexuality, Moslems who are permitted four wives and only have three, etc. With respect to this idealized cognitive model, a *bachelor* is simply an unmarried adult man.

This idealized model, however, does not fit the world very precisely. It is oversimplified in its background assumptions. There are some segments of society where the idealized model fits reasonably well, and when an unmarried adult man might well be called a bachelor. But the ICM does not fit the case of the pope or people abandoned in the jungle, like Tarzan. In such cases, unmarried adult males are certainly not representative members of the category of bachelors.

The theory of ICMs would account for such prototype effects of the category *bachelor* in the following way: An idealized cognitive model may fit one's understanding of the world either perfectly, very well, pretty well, somewhat well, pretty badly, badly, or not at all. If the ICM in which *bachelor* is defined fits a situation perfectly and the person referred to by the term is unequivocally an unmarried adult male, then he qualifies as a member of the category *bachelor.* The person referred to deviates from prototypical bachelorhood if either the ICM fails to fit the world perfectly or the person referred to deviates from being an unmarried adult male.

Under this account *bachelor* is not a graded category. It is an all-or-none concept relative to the appropriate ICM. The ICM characterizes representative bachelors. One kind of gradience arises from the degree to which the ungraded ICM fits our knowledge (or assumptions) about the world.

This account is irreducibly cognitive. It depends on being able to take two cognitive models—one for *bachelor* and one characterizing one's knowledge about an individual, say the pope—and compare them, noting the ways in which they overlap and the ways in which they differ. One needs the concept of "fitting" one's ICMs to one's understanding of a given situation and keeping track of the respects in which the fit is imperfect.

This kind of explanation cannot be given in a noncognitive theory—one in which a concept either fits the world as it is or not. The background conditions of the *bachelor* ICM rarely make a perfect seamless fit with the world as we know it. Still we can apply the concept with some degree of accuracy to situations where the background conditions don't quite mesh with our knowledge. And the worse the fit between the background conditions of the ICM and our knowledge, the less appropriate it is for us to apply the concept. The result is a gradience—a simple kind of prototype effect.

Lie

A case similar to Fillmore's *bachelor* example, but considerably more complex, has been discussed by Sweetser (1984). It is the category defined by the English word *lie.* Sweetser's analysis is based on experimental results by Coleman and Kay (1981) on the use of the verb *lie.* Coleman and Kay found that their informants did not appear to have necessary and sufficient conditions characterizing the meaning of *lie.* Instead they found a cluster of three conditions, no one of which was necessary and all of which varied in relative importance:

> A consistent pattern was found: falsity of belief is the most important element of the prototype of *lie*, intended deception the next most important element, and factual falsity is the least important. Informants fairly easily and reliably assign the word *lie* to reported speech acts in a more-or-less, rather than all-or-none, fashion, . . . [and] . . . informants agree fairly generally on the relative weights of the elements in the semantic prototype of *lie.*

Thus, there is agreement that if you steal something and then claim you didn't, that's a good example of a lie. A less representative example of a lie is when you tell the hostess "That was a great party!" when you were bored stiff. Or if you say something true but irrelevant, like "I'm going to

the candy store, Ma" when you're really going to the pool hall, but will be stopping by the candy store on the way.

An important anomaly did, however, turn up in the Coleman-Kay study. When informants were asked to define a *lie*, they consistently said it was a false statement, even though actual falsity turned out consistently to be the least important element by far in the cluster of conditions. Sweetser has observed that the theory of ICMs provides an elegant way out of this anomaly. She points out that, in most everyday language use, we take for granted an idealized cognitive model of social and linguistic interaction. Here is my revised and somewhat oversimplified version of the ICM Sweetser proposes:

THE MAXIM OF HELPFULNESS

People intend to help one another.

This is a version of Grice's cooperative principle.

THE ICM OF ORDINARY COMMUNICATION

(*a*) If people say something, they're intending to help if and only if they believe it.

(*b*) People intend to deceive if and only if they don't intend to help.

THE ICM OF JUSTIFIED BELIEF

(*c*) People have adequate reasons for their beliefs.

(*d*) What people have adequate reason to believe is true.

These two ICMs and the maxim of helpfulness govern a great deal of what we consider ordinary conversation, that is, conversation not constrained by special circumstances. For example, if I told you I just saw a mutual friend, under ordinary circumstances you'd probably assume I was being helpful, that I wasn't trying to deceive you, that I believed I had seen the friend, and that I did in fact see the friend. That is, unless you have reason to believe that the maxim of helpfulness is not applying or that one of these idealized models is not applicable, you would simply take them for granted.

These ICMs provide an explanation of why speakers will define a lie as a false statement, when falsity is by far the least important of the three factors discovered by the Kay-Coleman study. These two ICMs each have an internal logic and when they are taken together, they yield some interesting inferences. For example, it follows from (*c*) and (*d*) that if a person believes something, he has adequate reasons for his beliefs, and if he has adequate reasons for believing the proposition, then it is true. Thus, in the idealized world of these ICMs if *X* believes a proposition *P*, then *P* is true. Conversely, if *P* is false, then *X* doesn't believe *P*. Thus, falsity entails lack of belief.

In this idealized situation, falsity also entails an intent to deceive. As we have seen, falsity entails a lack of belief. By (*a*), someone who says something is intending to help if and only if he believes it. If he doesn't believe it, then he isn't intending to help. And by (*b*), someone who isn't intending to help in giving information is intending to deceive. Thus, in these ICMs, falsity entails both lack of belief and intent to deceive. Thus, from the definition of a lie as a false statement, the other properties of lying follow as consequences. Thus, the definition of *lie* does not need to list all these attributes. If *lie* is defined relative to these ICMs, then lack of belief and intent to deceive follow from falsity.

As Sweetser points out, the relative importance of these conditions is a consequence of their logical relations given these ICMs. Belief follows from a lack of intent to deceive and truth follows from belief. Truth is of the least concern since it is a consequence of the other conditions. Conversely, falsity is the most informative of the conditions in the idealized model, since falsity entails both intent to deceive and lack of belief. It is thus falsity that is the defining characteristic of a lie.

Sweetser's analysis provides both a simple, intuitive definition of *lie* and an explanation of all of the Coleman-Kay findings. The ICMs used are not made up just to account for *lie*. Rather they govern our everyday common sense reasoning. These results are possible because the ICMs have an internal logic. It is the *structure* of the ICMs that explains the Coleman-Kay findings.

Coleman and Kay discovered prototype effects for the category *lie*—situations where subjects gave uniform rankings of how good an example of a lie a given statement was. Sweetser's analysis explains these rankings on the basis of her ICM analysis, even though her ICM fits the classical theory! Nonprototypical cases are accounted for by imperfect fits of the lying ICM to knowledge about the situation at hand. For example, white lies and social lies occur in situations where condition (*b*) does not hold. A white lie is a case where deceit is not harmful, and a social lie is a case where deceit is helpful. In general, expressions such as *social lie, white lie, exaggeration, joke, kidding, oversimplification, tall tale, fiction, fib, mistake*, etc. can be accounted for in terms of systematic deviations from the above ICMs.

Although neither Sweetser nor anyone else has attempted to give a theory of complex concepts in terms of the theory of ICMs, it is worth considering what would be involved in doing so. As should be obvious, adjective-noun expressions like *social lie* do not work according to traditional theories. The category of social lies is not the intersection of the set of social things and the set of lies. The term *social* places one in a domain of experience characterized by an ICM that says that being polite is more

important than telling the truth. This conflicts with condition (*b*), that intent to deceive is not helpful, and it overrides this condition. Saying "That was a great party!" when you were bored stiff is a case where deception is helpful to all concerned. It is a prototypical social lie, though it is not a prototypical lie. The concept *social lie* is therefore represented by an ICM that overlaps in some respects with the lying ICM, but is different in an important way. The question that needs to be answered is whether the addition of the modifier *social* can account for this difference systematically. Any general account of complex concepts like *social lie* in terms of ICMs will have to indicate how the ICM evoked by *social* can cancel one condition of the ICM evoked by *lie*, while retaining the other conditions. An obvious suggestion would be that in conflicts between modifiers and heads, the modifiers win out. This would follow from the general cognitive principle that special cases take precedence over general cases.

Cluster Models: A Second Source of Prototype Effects

It commonly happens that a number of cognitive models combine to form a complex cluster that is psychologically more basic than the models taken individually. We will refer to these as *cluster models*.

Mother

An example is the concept *mother*. According to the classical theory, it should be possible to give clear necessary and sufficient conditions for *mother* that will fit all the cases and apply equally to all of them. Such a definition might be something like: *a woman who has given birth to a child*. But as we will see, no such definition will cover the full range of cases. *Mother* is a concept that is based on a complex model in which a number of individual cognitive models combine, forming a cluster model. The models in the cluster are:

- The birth model: The person who gives birth is the *mother*.

The birth model is usually accompanied by a genetic model, although since the development of egg and embryo implants, they do not always coincide.

- The genetic model: The female who contributes the genetic material is the *mother*.
- The nurturance model: The female adult who nurtures and raises a child is the *mother* of that child.
- The marital model: The wife of the father is the *mother*.
- The genealogical model: The closest female ancestor is the *mother*.

The concept *mother* normally involves a complex model in which all of these individual models combine to form a cluster model. There have always been divergences from this cluster; stepmothers have been around for a long time. But because of the complexities of modern life, the models in the cluster have come to diverge more and more. Still, many people feel the pressure to pick one model as being the right one, the one that "really" defines what a mother is. But although one might try to argue that only one of these characterizes the "real" concept of *mother*, the linguistic evidence does not bear this out. As the following sentences indicate, there is more than one criterion for "real" motherhood:

- I was adopted and I don't know who my real mother is.
- I am not a nurturant person, so I don't think I could ever be a real mother to any child.
- My real mother died when I was an embryo, and I was frozen and later implanted in the womb of the woman who gave birth to me.
- I had a genetic mother who contributed the egg that was planted in the womb of my real mother, who gave birth to me and raised me.
- By genetic engineering, the genes in the egg my father's sperm fertilized were spliced together from genes in the eggs of twenty different women. I wouldn't call any of them my real mother. My real mother is the woman who bore and raised me, even though I don't have any single genetic mother.

In short, more than one of these models contributes to the characterization of a *real mother*, and any one of them may be absent from such a characterization. Still, the very idea that there is such a thing as a *real mother* seems to require a choice among models where they diverge. It would be bizarre for someone to say:

- I have four real mothers: the woman who contributed my genes, the woman who gave birth to me, the woman who raised me, and my father's current wife.

When the cluster of models that jointly characterize a concept diverge, there is still a strong pull to view one as the most important. This is reflected in the institution of dictionaries. Each dictionary, by historical convention, must list a primary meaning when a word has more than one. Not surprisingly, the human beings who write dictionaries vary in their choices. Dr. Johnson chose the birth model as primary, and many of the applied linguists who work for the publishers of dictionaries, as is so often the case, have simply played it safe and copied him. But not all. *Funk and Wagnall's Standard* chose the nurturance model as primary, while the *American College Dictionary* chose the genealogical model. Though

choices made by dictionary-makers are of no scientific importance, they do reflect the fact that, even among people who construct definitions for a living, there is no single, generally accepted cognitive model for such a common concept as "mother."

When the situation is such that the models for *mother* do not pick out a single individual, we get compound expressions like *stepmother, surrogate mother, adoptive mother, foster mother, biological mother, donor mother*, etc. Such compounds, of course, do not represent simple subcategories, that is, kinds of ordinary mothers. Rather, they describe cases where there is a lack of convergence of the various models.

And, not surprisingly, different models are used as the basis of different extended senses of *mother*. For example, the birth model is the basis of the metaphorical sense in

– Necessity is the mother of invention.

while the nurturance model is basis for the derived verb in

– He wants his girlfriend to mother him.

The genealogical model is the basis for the metaphorical extension of *mother* and *daughter* used in the description of the tree diagrams that linguists use to describe sentence structure. If node *A* is immediately above node *B* in a tree, *A* is called the *mother* and *B*, the *daughter*. Even in the case of metaphorical extensions, there is no single privileged model for *mother* on which the extensions are based. This accords with the evidence cited above which indicates that the concept *mother* is defined by a cluster model.

This phenomenon is beyond the scope of the classical theory. The concept *mother* is not clearly defined, once and for all, in terms of common necessary and sufficient conditions. There need be no necessary and sufficient conditions for motherhood shared by normal biological mothers, donor mothers (who donate an egg), surrogate mothers (who bear the child, but may not have donated the egg), adoptive mothers, unwed mothers who give their children up for adoption, and stepmothers. They are all mothers by virtue of their relation to the ideal case, where the models converge. That ideal case is one of the many kinds of cases that give rise to prototype effects.

CHAPTER 5

Metonymic Models

Metonymy is one of the basic characteristics of cognition. It is extremely common for people to take one well-understood or easy-to-perceive aspect of something and use it to stand either for the thing as a whole or for some other aspect or part of it. The best-known cases are those like the following:

– One waitress says to another, "The ham sandwich just spilled beer all over himself."

Here *the ham sandwich* is standing for the person eating the sandwich. Another well-known example is the slogan:

– Don't let El Salvador become another Vietnam.

Here the place is standing for the events that occurred at that place. As Lakoff and Johnson (1980, chap. 8) showed, such examples are instances of general principles; they do not just occur one by one. For example, English has a general principle by which a place may stand for an institution located at that place:

– The White House isn't saying anything.
– Washington is insensitive to the needs of ordinary people.
– The Kremlin threated to boycott the next round of talks.
– Paris is introducing shorter skirts this season.
– Hollywood isn't what it used to be.
– Wall Street is in a panic.

In each of these cases, a place like *The Kremlin* is standing for an institution located at that place, like the Soviet government. Moreover, the principle applies to an open-ended class of cases, not to any fixed list. For example, suppose that I am running a company that has many branch offices, including one in Cleveland, and I have asked each branch to send