# Knowledge Representation

Arthur B. Markman
*University of Texas at Austin*

LAWRENCE ERLBAUM ASSOCIATES, PUBLISHERS
1999 Mahwah, New Jersey London

Chapter 1

# Foundations

In the study of the mind, cognitive scientists seek explanations of mental life, which consists of perceptions, emotions, social interactions, and cognitive abilities. In many of these explanations, they refer to goals, beliefs, mental images, concepts, and other mental entities. They are also concerned with retrieval, analogy, inferencing, reasoning, categorization, and many other processes that create, combine, and use the information "in our heads." The aim of this book is to consider ways of thinking about goals, beliefs, mental images, concepts, and other mental entities to understand how different decisions about the way to characterize these entities affects what is easy to do with them and what is hard to do with them.

In particular, this book is concerned with the question of mental representation. That is, what formats are used for the information that makes up mental life (and how is the information used)? In this book, I explore a variety of options for representing information and focus mainly on the assumptions made by different representational formats and on the ways these assumptions affect what is easy or hard to do with them. Before the discussion can proceed, however, several preliminary issues must be dealt with. First, what is a representation? Second, why worry about the nature of mental representations? Finally, how do representations fit into the study of cognition? Chapter 1 addresses these topics.

## AN EXAMPLE

The issue of mental representation may seem uninteresting. Perhaps there are not many options for representing a situation, or the choice of representation may be irrelevant to what a model of mental processing can explain. Even if there are differences between models, these differences may have no practical significance for the way psychology is carried out as a science. In this section, I present an example demonstrating that there are often many different ways that something can be represented, that differences in representations do affect the explanatory capability of a model, and that the choice of representations has important implications for how psychology is done.

My example comes from the study of people's ability to do logical reasoning. The prototypical version of the task was presented by Wason and Johnson-Laird (1972) and has become known as the Wason selection task. In this task, researchers show people four cards on a table and tell them that all the cards have a letter on one side and a number on the other. The four cards are laid out so that they face the subject as shown in Figure 1.1. Then, the subject is asked to point to the smallest number of cards necessary to test the truth of the rule "If there is a vowel on one side of the card, then there is an odd number on the other side of the card."

Countless researchers (Johnson-Laird, 1983; Rips, 1994) have examined variations of this task. When the problem is framed as presented in Figure 1.1, people often have difficulty getting the right answer. Most people will say that the card with the letter *A* must be turned over. Few people think that they have to turn over the card with the letter *J*. People are split on what to do with the numbers. Some think that both numbers can be ignored, some feel the *seven* must be turned over, some feel the *four* must be turned over, and some feel that both cards must be turned over. The correct answer is that the *A* and the *four* must be turned over: If there is an even number on the other side of the *A* card, the rule is invalid, and if there is a vowel on the other side of the *four* card, the rule is invalid. The *J* need not be turned over; the rule does not apply to it, and it does not matter what is on the other side. The *seven* need not be turned over;

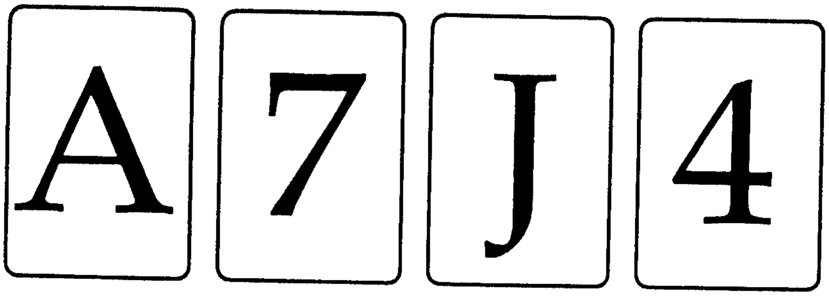

FIG. 1.1. Wason selection task.

if there is a vowel on the other side, the rule applies and is valid, but if there is a consonant on the other side, the rule simply does not apply. Thus, turning over the *seven* does not provide a way to invalidate the rule.

How can I explain subjects' difficulty with this task? Perhaps people represent the task as one of logical reasoning. Turning over the *A* card corresponds to the valid logical inference schema *modus ponens*:

$$\begin{array}{l} \text{IF } P\text{, then } Q \\ P \\ \hline Q. \end{array} \tag{1.1}$$

This schema reads "If some statement *P* is true, then some statement *Q* is true. Statement *P* is true. Therefore, Statement *Q* is true." For this schema, any statement that can be either true or false may play the roles of *P* and *Q*. Turning over the *four* card requires the logical schema *modus tollens*:

$$\begin{array}{l} \text{If } P\text{, then } Q \\ \text{NOT } Q \\ \hline \text{NOT } P. \end{array} \tag{1.2}$$

Not all schemas give rise to valid rules of inference. The valid rules are those for which if the premises (i.e., the statements above the line) are true, then the conclusion (i.e., the statement below the line) is guaranteed to be true. One example of an invalid schema is *affirming the consequent*:

$$\begin{array}{l} \text{IF } P\text{, then } Q \\ Q \\ \hline P. \end{array} \tag{1.3}$$

The problem with this schema is that it fails to take into account that the statement *Q* can be the case for some reason other than the rule "If *P*, then *Q*." This schema would be valid if the rule were "*Q* is true if *and only if* *P* is true" (sometimes written IFF *P*, then *Q*).

Taking logical rules seriously as a representation of people's ways of reasoning suggests that correct performance on the Wason selection task requires both *modus ponens* and *modus tollens*, but not incorrect schemas like *affirming the consequent*. Because most people turn over the *A* card and fewer people turn over the *four* card, *modus ponens* must be an easier rule to learn than is *modus tollens*. Accounts of logical reasoning of this type have been proposed by Rips (1994) and Braine, Reiser, and Rumain (1984). This account of reasoning assumes that people use general rules of reasoning across domains. By adopting this framework, a researcher makes certain questions particularly interesting to answer. For example, a researcher who assumes this representation may focus on the rules people tend to have, the factors that promote the acquisition of new rules, and the factors that control whether people recognize that a particular rule is relevant in a given context.

According to an alternative account of logical reasoning ability, however, people do not have logical rules that apply across domains. After all, logical rules do not care about the content of the statements *P* and *Q*. As long as a situation has the right form, the logical rules apply. According to one such account, people have a *mental model* of a situation about which they are going to reason (see chap. 9). Mental models are not general schemas of inference but instantiations of particular situations. This account suggests that problems framed in an abstract way (like the Wason selection task) are difficult, because it is difficult to construct models for abstract situations. Thus, a problem with the same structure (an isomorphic problem) may be easier to solve if it is in a domain for which it is easy to construct a model.

This view of representation suggests that the selection task should be tried with different problem contents. As an example, imagine you are working for the security patrol of a college on a Saturday night, and it is your job to make sure that campus bars serve alcohol only to people of the legal drinking age (21 years old in the United States). You enter a bar and see one person you know to be 18 years old, a second you know to be 22 years old, a third person, whom you do not know, holding a beer, and a fourth, unknown to you, drinking club soda. Which people must you check to ensure that the bar is satisfying the rule "If a person has a drink, then he or she is over 21"? College undergraduates given versions of the selection task in familiar domains like this performed quite well. Nearly all knew that only the 18-year-old and the person drinking beer need to be checked. This problem is isomorphic to the task with the cards, but people have much less difficulty with the concrete version (see Johnson-Laird, Legrenzi, & Legrenzi, 1972).

Mental models are not exactly the same as logical rules. Although mental models can be described as having rules, the scope of these rules differs from that of logic. With a particular logical rule, anything with the proper form can be reasoned about. In contrast, the procedures for constructing mental models are domain specific. A person may have rules for reasoning about drinking in bars without having rules for reasoning about genetics or abstract logical forms. For those who have adopted a framework based on logical rules, the content effects discovered in the selection task are difficult to explain. Rips (1994) argued that content effects in this task may reflect people's remembering what happened in their own personal experience and that this personal experience augments but does not replace logical rules. For example, when given the selection task in the context of verifying the rule about the legal age for drinking, people may just recall a situation in which they were in a bar and remember who was asked for identification. In this case, no rules were used at all; the answer to the problem was just remembered. Assuming that reasoning uses logical rules of inference makes it easy to explain logical reasoning abilities at the expense of making content effects more difficult to explain.

People's performance on a psychological task may often be explained in many ways, each of which has a different approach to mental representation. Each way may provide a good account of the phenomenon being studied, but the approaches may differ in their predictions for subsequent studies that should be designed and carried out. Indeed, as I discuss next, adopting particular representational assumptions affects which new questions are most interesting to answer.

## WHAT IS A REPRESENTATION?

Mental representation is a critical part of psychological explanation, but it has also been a source of great confusion. Different researchers have used the word *representation* in different ways. Psychologists have used representation in somewhat different ways from other cognitive scientists, such as philosophers and computer scientists, who are interested in representation. To avoid confusion, I offer a broad definition of representation, one that includes all things that cognitive scientists have considered representations, although it may admit some things that people may feel uncomfortable calling representations, or at least uncomfortable thinking of as psychological representations. My definition of representation has four components. The first two components of representation are:

1. **A represented world:** the domain that the representations are about. The represented world may be the world outside the cognitive system or some other set of representations inside the system. That is, one set of representations can be about another set of representations.
2. **A representing world:** the domain that contains the representations. (The terms *represented world* and *representing world* come from a classic paper by Palmer [1978a].)

As an example, consider various representations of the items pictured in the top row of Figure 1.2. These items are the represented world for this example. In this world, there are three objects of interest, an ice cube, a glass of water, and a pot of water on a fire. I can choose to represent many aspects of this world, but for now, I focus on the temperature of the water. This representational decision has consequences. If I represent only

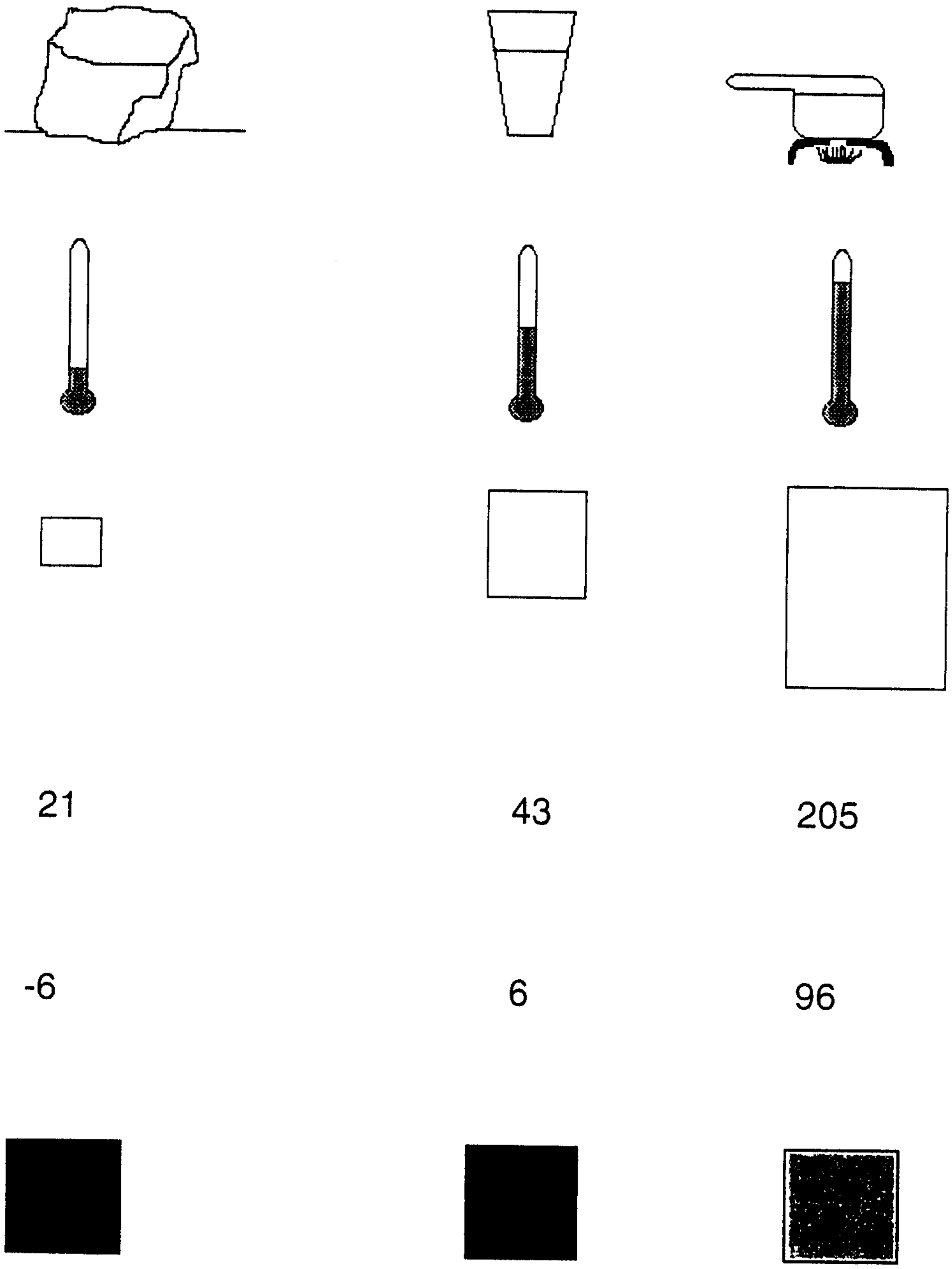

FIG. 1.2. Various ways of representing temperature. The top row depicts water that is frozen, at room temperature, and boiling. The next two rows depict possible analog representations. The two following rows show numerical temperature notations. Finally, the last row depicts temperature with the darkness of the square.

the temperature of the water, all the rest of the information about the situation is lost, including the shape of the ice cube, the size of the glass of water, and the degree of curvature of the handle of the pot. This point is not trivial: In all known representational systems, the representing world loses information about the represented world.

In modern culture, the representation of temperature, as in the second row of Figure 1.2, often appears as the height of mercury in a thermometer

That is, I can use the height of mercury as a representing world, in which the higher the line of mercury, the greater the temperature. In this representation, however, a few important issues lie buried. First, the height of mercury in a thermometer works as a representation of temperature, because there is a set of rules that determine how the representing world corresponds to the represented world. Thus, the third component of my definition of representation is:

3. **Representing rules:** The representing world is related to the represented world through a set of rules that map elements of the represented world to elements in the representing world. If every element in the represented world is represented by a unique element in the representing world, there is an *isomorphism* between the represented and representing worlds. If two or more elements in the represented world are represented by one element in the representing world, there is a *homomorphism* between the represented and the representing worlds.[1]

As an illustration of this component of the definition, when temperature is represented as the height of mercury in a thermometer, each temperature is reflected by a unique height of mercury. The specific height that the mercury reaches is determined by the circumference of the thermometer as well as by the physical laws that govern the expansion of mercury with changes in temperature. Because each temperature has its own unique height, there is an isomorphism between the temperature in the represented world and the height of mercury in the representing world, but not all representations of temperature need to be isomorphisms. If a digital thermometer that gave readings on the Fahrenheit temperature scale (as in the second row of Figure 1.2) gave readings accurate to only 1 degree, any temperature between, say, 20.5 degrees and 21.4 degrees would be represented as 21 degrees. In this case, the relation between the represented and representing worlds is a homomorphism. When there is a homomorphism between the representing and represented worlds, the representation has lost information about what it is representing.

Another issue that arises with this example is that nothing inherent in a mercury thermometer alone makes it a representation. Since the dawn of time (or soon thereafter), mercury has had the property of expanding and contracting with changes in temperature, but mercury was not always a representation of temperature. For something to be a representation,

[1] There is much debate in philosophy about how physical systems (like minds) have the power to represent things in the external world, but in this book, I am not concerned with solving this problem. Instead, I assume that cognitive systems have the capacity for representation, and I focus on proposals in cognitive science for the nature of these representations.

some process must use the representation for some purpose. In this culture, having been schooled in the use of a thermometer, people can use the column of mercury as a representation of temperature. A vervet monkey who lacks the mathematical skills and cultural upbringing (among other things) to read a thermometer cannot use the column of mercury as a representation of temperature. More broadly, something is a representation only if a process can be used to interpret that representation. In this case, the combination of the thermometer and the person who can read it makes the thermometer a representation. More generally, the fourth component of a representation is:

4. **A process that uses the representation:** It makes no sense to talk about representations in the absence of processes. The combination of the first three components (a represented world, a representing world, and a set of representing rules) creates merely the potential for representation. Only when there is also a process that uses the representation does the system actually represent, and the capabilities of a system are defined only when there is both a representation and a process.

The importance of processes when thinking about representations cannot be underestimated (J. R. Anderson, 1978; Palmer, 1978a). In the temperature example, there is no representation until someone can use the thermometer to read off the temperature. In general, it may seem obvious that certain cognitive processes can be explained by a representation, but in many instances two very different kinds of representations can make exactly the same predictions when the right set of processes acts over them.

To demonstrate how the four components of representation interact to create a representation, we return to the temperature example. The rectangles in the third row of Figure 1.2 can also be representations of temperature. For example, the area of each rectangle could be used as a representation of a particular temperature. In this case, comparing two temperatures may involve laying one rectangle on top of another to see which is larger: A larger rectangle corresponds to a higher temperature. Of course, a different set of representing rules and processes completely changes the interpretation of this representation: If the heights of the rectangles are used to represent temperature, pairs of temperatures can be compared by laying the rectangles next to each other. It is easy to generate other possibilities: For example, smaller rectangles can represent higher temperatures. For each possibility, the representing rules and associated processes for interpreting the representation must be configured accordingly.

The representations in the second and third rows of Figure 1.2 depict a continuous quantity with another continuous quantity. Once the length of the line is linked by a representing rule with the temperature of the object, a change in the length of the line can be interpreted as a change in temperature. Using one dimensional quantity to represent another seems to provide some information for free. These representations are often called *analog*, because the representing world has an inherent structure that governs how it operates and the relations between aspects in the representing world are not arbitrary. For example, it is a fact about spaces that if line *A* is longer than line *B* and line *B* is longer than line *C*, line *A* is also longer than line *C*. Length is an appropriate representation for temperature, because temperature has the same transitive structure as length. If temperature did not have a transitive structure, length would not be an appropriate representing world to use for temperature.

Not all representations are analog. The fourth row of Figure 1.2 shows such a representation. In modern culture, people use numerical representations of quantities such as temperature all the time. This representation is very different from those in the second and third rows: Making the numbers taller or shorter does not signal changing the temperature of the objects; only changes in the digits change the representation. Nothing inherent in the scratches of ink requires the number 21 to be larger than the number 20 or smaller than the number 22. Rather, a system of representing rules links the written numerals to the represented world of abstract mathematical quantities. These representations are often called *symbolic* because a convention is established to link all the elements in the representing world. The relation among elements in the representing world is arbitrary and could have occurred in some other way had the representing rules been differently constructed. (The arbitrary, conventionally established use of symbols in mental representation is similar to the use that, for example, allows the symbol Σ to play one role in Greek writing and a different role in mathematical equations.)

For a symbol system like Arabic numerals to be used as the basis of a representation of the represented world of temperature, a set of representing rules must be established between symbols and temperature. First, there must be rules that map the numerals onto numbers, but even after this mapping has been established, there are many possible ways to map the numbers onto temperatures. For example, with the Fahrenheit temperature scale as representing rules, there is a set of correspondences between the represented and representing worlds different from that with the Celsius scale. An infinite number of sets of representing rules can be constructed to map numbers onto temperatures. The same point is true for analog representations (such as using lines to represent temperatures): Different lengths of a line can represent the same degree of temperature change. The particular correspondence between temperature and line length is established by the representing rules.

A particular representation makes some information obvious and other information difficult to extract (Marr, 1982). The length of a column of mercury as a representation of temperature makes it easy to make direct comparisons between pairs of temperatures. A simple procedure of laying two lines next to each other and seeing which extends further accomplishes this task. To compare two numbers, in contrast, extensive knowledge of the system of symbols underlying numbers and an understanding of numerical relationships, must be brought to bear. Not all things are easier to do with the length representation of temperature, though. If a specific value for temperature is required to make a complex calculation, say for understanding a chemical reaction, the length representation is poorly suited as a representation of temperature; for this purpose, the numerical representation may be better.

To summarize, representations have four components. At the heart of a representation is a representing world that is used to represent information in the represented world. The particular representations in this world are bound to the represented world by representing rules that relate aspects of the representing world to aspects of the represented world. The process of representing some world typically produces a loss of information, because information can be used only when there is a procedure for extracting it. I have made a distinction between *analog* representations, for which the relations among elements in the represented world are fixed by the structure of the representational system, and *symbolic* representations, for which the relations among elements in the represented world are arbitrary and must be fixed by convention. Finally, any representational choice makes some information easy to find but may make other information very difficult to determine.

## Representations and Cognitive Representations

My working definition of representation is quite broad. For example, according to this definition, a thermostat has representations although a thermostat is not a cognitive system. What exactly constitutes a cognitive system or cognitive representation is a difficult question that has occupied researchers in cognitive science for some time. No satisfactory definition exists that includes all and only things that all researchers are happy calling representations, but I hope to demonstrate the range of things that are good candidates for psychological representations. In this way, I can triangulate on a good definition for cognitive representation.

Before we examine the notion of cognitive representations in more detail, however, there is one danger with defining a cognitive system that we must discuss explicitly. There is a strong intuition that a thermostat is not a cognitive system. After all, a thermostat is not that interesting a device. The bimetallic plate changes its shape with the temperature in the room, and at some point the change causes a switch to close and to turn on the heat (or perhaps the air conditioning). Later, the change in temperature in the opposite direction causes the switch to open, and the heat (or air conditioning) stops. Why is this not cognitive?

J. A. Fodor (1986) gave an answer to this question. He argued that the behavior of systems like thermostats is well described, perhaps even best described, by using the principles of physics and chemistry. A bimetallic plate changes its shape because the two metals expand at different rates (a change predicted by laws of physics and chemistry). Thus, although a thermostat has a representation, it is a representation that needs no principles of psychology to be understood, and thus it is not a cognitive representation.[2] This way of distinguishing between cognitive and noncognitive representations seems reasonable. Although the behavior of any representational system can be described by the laws of physics at some level, no interesting generalizations from physics or chemistry can explain how a cognitive system (like a brain or a suitably programmed computer) represents information, and so it is necessary to appeal to other sciences.

An inappropriate answer (in my view) to the question of what makes something a cognitive representation is that a thermostat is a deterministic device. The physics of thermostats is well enough understood to predict their behavior with striking accuracy. A complete psychology may allow predictions of the behavior of humans and other animals with alarming effectiveness as well. I raise this point here, however, because implicit in many discussions of representation is the notion that a cognitive system has an element of free will in it. I do not choose a definition of cognitive system in a way that assumes that there is (or is not) free will.

### The Meaning in a Representation

In a cognitive representation, the representation is an internal state; that is, in humans, mental representations are in the head. The nature of the represented world is controversial. Is the represented world in the head, outside the head, or some combination of the two? Cognitive scientists have often assumed that some represented worlds are outside the head and others are inside. In order to look more at what it means for a representation to be about something, we must explore some work in the philosophy of mind.

[2] Actually, Fodor argued that thermostats do not have representations at all and that what makes something a representation is having properties that cannot be explained by the laws of basic sciences (what he called *non-nomic* properties). This position does not explain how a thermostat makes contact with its environment to do something interesting in a way that a rock heating in the sun does not. See Markman and Dietrich (1998) for a more complete discussion of Fodor's view and problems with it.

Philosophers have been concerned with the notion of *intentionality* (Dennett, 1987; Dietrich, 1994; Searle, 1992). Rather than referring to the familiar idea that something may be done with intent or on purpose, the philosophical concept of intentionality refers to what a representation is about. For example, my representation of the computer screen I am looking at is about this computer screen. My belief that chocolate ice cream is good is about chocolate ice cream. My belief that unicorns do not exist is about unicorns. Defining *aboutness,* however, is not straightforward. For my belief about the computer screen, it is enough that there is a computer screen in front of me (barring a very convincing hallucination). Likewise, my belief about chocolate ice cream can refer to instances of chocolate ice cream in the world, particularly those instances I have experienced in the past (that were good). Unicorns are more problematic: There are none, and never were. Thus, it is not enough to assume that representations are about things in the world because not everything represented is in the world (or ever was). Some things may be abstract concepts without good visualizable forms. Finally, even when a representation *is* about something in the outside world, there may be a mismatch between what is in the world and the way I represent it. On a dark foggy night, I may represent something as a black cat, only to find out too late that it is a skunk. A theory of representation must allow such mistakes to occur.

A complete catalog of theories of intentionality in philosophy would take up more room than I have in this chapter (or in this book; see J. A. Fodor, 1981; J. A. Fodor & Lepore, 1992; Stich & Warfield, 1994). To understand representation, it is important to think about how the elements in a representation can mean something. One solution to this problem (discussed again in later chapters) is conceptual role semantics, in which the meaning of a representational element is fixed by its relations to other representational elements. This situation is analogous to a dictionary, in which a word is defined in terms of other words. For example, the glossary of an introductory psychology textbook may define the term *olfaction* as "the sense of smell." This definition is helpful only if there already is a meaning for the phrase "the sense of smell" (and you know what that definition is).

A conceptual role semantics has two problems: First, the meanings of at least some elements in the representation must be known, or none of the elements means anything. The representational elements with known meanings are the *grounded* elements. Without knowing the meanings of any words, it is not helpful to look up words in a dictionary: Everything is gibberish in this case. In the chapters that follow, it is worth thinking about which elements in the representing world may be grounded, and how may the grounding take place?

The second problem with conceptual role semantics is holism (J. A. Fodor & Lepore, 1992). If representational elements are given meaning by their relations to other representational elements, the meaning of any one element depends on every other representational element. According to this view, two people's concepts of *dog* differ because each knows different things about dogs, and also about the 1986 New York Giants. If the meaning of any concept depends on the meaning of every other concept, then how can people function without accessing all information at all times? If each person's concepts differ from every other person's concepts, because of differences in past experience, communication is impossible: One person's meanings of the concepts used in a discourse must differ radically from another person's concepts on the basis of differences in past experience. The holism problem requires that cognitive systems be able to do some processing without having to make use of every piece of their knowledge for every process. Again, in the chapters that follow, it is worth thinking about how particular representations avoid having to access and use every piece of known information to function.

A final problem that philosophers have often raised in conjunction with discussions about representation concerns how representations are interpreted. If I have a picture of the Grand Canyon, I believe the picture represents the Grand Canyon because of particular color saturation patterns that map onto color saturations that were at the Grand Canyon at the time the picture was taken. When I look at it, because I have the right kind of visual system, I can interpret the picture and extract information from the representation. The problem comes with thinking about cognitive representations. The representing world in a cognitive representation is assumed to be internal to the organism. Who looks at the representation to interpret it? There cannot be another person in my head (a *homunculus*) who looks at my representations, because then who would interpret the representations in the homunculus's head?

Cognitive scientists have generally avoided this conundrum by assuming that the cognitive system is a computational device. That is, the cognitive system has representations, and it also has processes that manipulate the information in these representations, just as a familiar digital computer can have data structures, which can be manipulated by procedures in computer programs. Digital computers are able to carry out algorithms, because they have instructions encoded in them to allow them to follow a program in the same way that a cook follows a recipe.[3]

[3] The ability to follow a program is based on the theoretical concept of a Turing machine. A description of Turing machines is beyond the scope of this chapter; interested readers should consult the description of Turing machines by Johnson-Laird (1988). A clever introduction to Turing machines appeared in Barwise and Etchemendy's work (1993b); they provided a computer program that allows readers to construct Turing machines to solve a variety of problems.

In this section, I have raised two important philosophical issues about representation. The first is intentionality (i.e., what a representation is about): How is the representing world connected to the represented world? The second is computation: There is a danger when positing psychological theories of requiring an intelligent agent to interpret the representations in it. According to the concept of computability derived from Turing machines, a process designed to make use of a representation can be carried out without needing such an intelligent agent.

## THREE DIMENSIONS OF VARIATION IN REPRESENTATIONS

How does one representational format differ from another? Are the differences merely a matter of notation, or do actual substantive issues separate the types of representations? Proposals for representations can vary along many dimensions (see also Markman & Dietrich, 1998).[4] As a demonstration that these dimensions of variation are substantive, I consider three: the duration of representational states, the presence of discrete symbols, and the abstractness of representations. In the following section, I discuss some general criteria for deciding that one proposal for representation is better than another.

The first dimension of variation is in the duration of representational states. The definition of representation given here does not require that representational states exist for any particular time. In the case of a mercury thermometer, representational states are instantaneous; the height of the mercury in the thermometer represents the temperature at the moment. Any changes in temperature change the height of the mercury and leave the system without any memory of past states. Representations may also endure for long periods. I can remember the day that my parents and I moved from an apartment to a house when I was about 3 years old (some 28 years before I am writing this). The fact that I have a mental image of the moving truck behind our car means that some representation of this event has endured in my cognitive system for a long time (although my current mental image of this state may reflect only a transient activation of neurons in my brain). Thus, different representational systems may focus on transient or enduring representational states.

A second important dimension of variation is the presence of discrete symbols. Many representational formats assume that discrete elements in the representing world bear some relation to elements in the represented

[4] Markman and Dietrich (1988) actually discussed five dimensions of variation, but three are most central for this discussion.

world. When the relation between these discrete elements and the things they represent in the represented world is arbitrary, these discrete elements are called symbols. Although symbols are common in representational systems (see chaps. 3–9), they are not obligatory. For example, as I discuss in chapter 2, many systems use space as a representation. Space is continuous and hence does not divide the representing world into discrete parts. Thus, symbols are commonly used in representations but are not required.

The issues of duration and symbol use are not trivial and have been the source of some controversy in cognitive science. Indeed, theorists who have focused on representations that exist for only short periods and do not require explicit symbols have considered the possibility that cognitive systems have no representations at all. One example presented by van Gelder (1992; see also Thelen & Smith, 1994) involves Watt's apparatus used as the governor for a steam engine. The mechanism, shown in Figure 1.3, spins around; the faster it spins, the higher the balls on the outside rise. As the balls rise, they close a valve that lets steam flow through the engine; this process reduces the pressure and causes the mechanism to spin more slowly. The decrease in the rate of spin lowers the balls, which causes the valve to open more, thereby increasing the pressure, and so on. This elegant machine keeps steam engines from exploding by keeping pressure in the engine from rising too high.

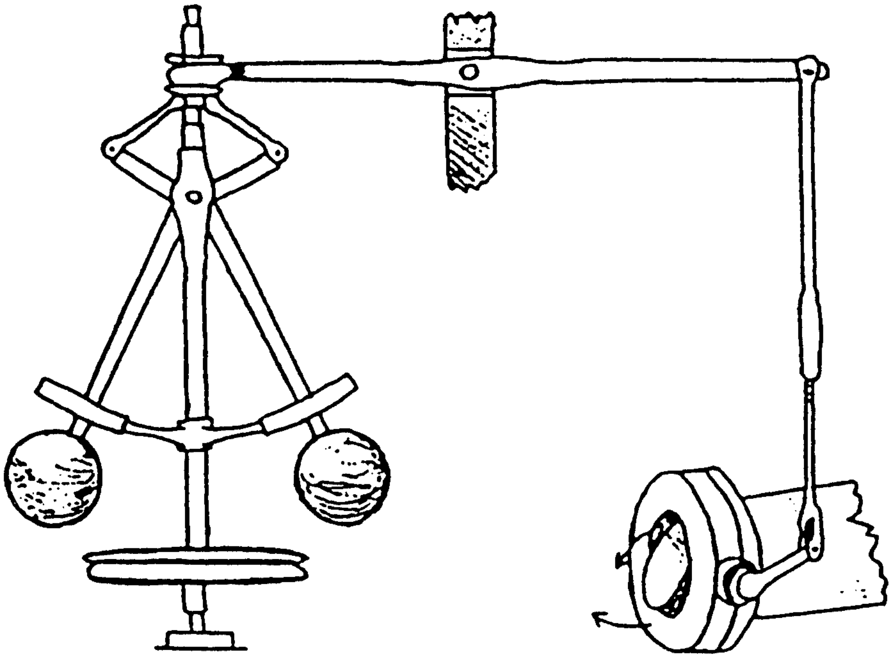

FIG. 1.3. Diagram of Watt's steam engine governor. From *Dynamic Systems Approach*, by E. Thelen and L. B. Smith (1994). Copyright © 1994 by MIT Press. Reprinted with permission.

Van Gelder suggested that this mechanism requires no representation. The function of the governor is clearly not carried out in the way that it would be if one programmed a digital computer to reason qualitatively about a system (see chap. 9). There is no mental model of the inside of the steam engine, and nothing in the system allows the governor to reason about aspects of the engine in case of a problem. Furthermore, there is no explicit symbolic representation of pressure that a problem-solving program can use to reason about the optimal level of pressure needed to run the engine at peak efficiency. Indeed, if the valve jammed in its open state, the governor would continue to spin faster as the steam built up until the engine exploded despite the governor's best efforts.

Although the governor has no symbolic representation of the pressure inside the steam engine, it does represent the pressure through the speed at which the governor spins. This speed, which is a stable representation of pressure, is a function of the weight of the balls on the sides of the apparatus. There is also a process that extracts information from the representation, namely the lever connecting the arms of the governor to the valve. This representation is an excellent example of how a particular representation may make some information obvious at the expense of other information. The system can react only to current pressure; it cannot reason about the information in any other way.

This representation is also not enduring: At any moment, all that is available is a measure of the current pressure inside the engine. There is no way to compare the current pressure to the pressure at any previous time, because this information is not stored. The advantage of this representational choice is that it provides a simple, elegant method that allows the governor to carry out its function. If it was important for the governor to make comparisons with past states, however, this representation would be insufficient: The system would need a representational format that was more enduring. Thus, the governor uses simple, nonenduring representations to carry out its task. Like the thermostat described earlier, however, the representations in the governor are well described by physical laws, and so they are not good candidates to be cognitive representations.

Finally, consider the issue of abstractness. Since Aristotle, human thought has been prized for its logical facility. Unlike any other species, people can reason in ways that are independent of the content of a domain and can see similarities across items that appear to be wildly dissimilar on the surface. The recognition that human cognition is privileged in its ability to reason abstractly has led to the supposition that cognitive representations seek abstractness. As a result, many proposals for mental representations have focused on how to represent cases that are abstracted across surface details.

More recently, as exemplified by the content effects in logical reasoning problems discussed earlier, psychologists have begun to study the ways in which reasoning is embedded in the content of domains. The move away from abstract representations does not remove representations from the picture in cognitive science but simply stresses that in models of psychological processing more abstract does not necessarily mean better. This abstraction process has also taken place at the level of theory. Studies of higher cognitive processing have often assumed that mental representations used by higher cognitive processes are independent of the representations used by perceptual systems that take in information from the world and independent of the representations used by motor systems to act on the world. This assumption may be unjustified. (I discuss the importance of domain-specific information in cognitive processing in chaps. 8 and 9.)

A similar argument can be made about the temporal duration of representational states. When cognition is viewed as a process that aspires to a logical ideal, creating enduring representations that exist across the life of an organism is the ideal state. Indeed, cultural systems may be one way that humans try to build lasting representational structures in different individuals over time. There is, however, much flexibility in cognitive processing. The Necker cube (Figure 1.4) is one simple example. If a person stares at the cube long enough, perception flips back and forth between two stable states that place two different sides at the front of a three-dimensional cube. Because the two interpretations of the cube are incompatible, people do not consciously perceive both interpretations at the same time. In this case, the representation of the cube changes from one state to another every few seconds. These changes in the psychological construal of different situations can probably occur for conceptual situations as well as for perceptual ones like this example. This may occur when people go back and forth while making a decision, first favoring one option, and then the other. Repre-

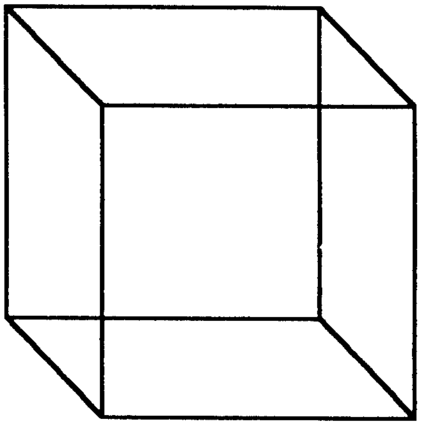

FIG. 1.4. Necker cube.

sentational structures, however, are not constructed once to remain unchanged in a psychological system. Rather, some psychological processes require representations that change rapidly over time.

The view that human beings aspire to a logical ideal carries with it an assumption that the optimal form of psychological representation is content free and enduring. Many proposals about psychological representations carry these assumptions by default, but they are not obligatory. Demonstrations that psychological representations contain information about specific domains or that representations change dynamically during the course of processing are simply more data points that constrain the way people think about representation. In this book, I have made an effort to construe representation broadly, because there is little hope of understanding the effects of content and temporal change in psychology without thinking carefully about representation.

## EVALUATING PROPOSALS FOR REPRESENTATION

I explore different proposals for mental representation in this book. These proposals differ along a number of lines including the three just described (i.e., the enduringness of the representation, the presence of symbols, and the abstractness of the representation). In view of these proposals, which representation is best for explaining a given cognitive process?

One important criterion for selecting a representation is its power. Not all representations are powerful enough to represent all represented worlds. For example, in chapter 5, I discuss representations that permit *quantification* (i.e., representations that allow people to say things like "Some dogs are small"). Some representations cannot represent quantified assertions (i.e., they can talk about dogs and even small dogs, but they cannot represent that *some* dogs are small). Thus, if a particular model requires the ability to handle quantified statements, then a representational system that is not powerful enough to represent quantified statements is not appropriate. A representation may be rejected as a candidate for a particular process because it is not expressive enough to represent the information needed for that domain.

A person faced with two or more representations that are powerful enough for the domain can find the problem of deciding between them quite difficult. As I have discussed, processes that use representations are an important part of the concept of representation. J. R. Anderson (1978) demonstrated that if two different representing worlds are both sufficient to represent the information in a represented world, it is often possible to find in each domain a set of processing assumptions that allow systems using either representation to explain the same set of data. Thus, it is generally impossible to argue decisively against a particular form of mental representation on the basis of data from a set of studies.

Because of this difficulty, there may seem to be no point in trying to determine which model of representation is best; the question is not answerable. There are, however, other criteria for favoring one model of representation over another. One critical use of representation is in generating further research questions. As one example, consider behavior that is rule governed. Traditional symbolic accounts of phenomena in developmental psychology have led researchers to focus on stages of development: Piaget constructed a theory about the levels of cognitive competence that children achieve and the ages at which they achieve these levels of performance. He then described general mechanisms that allow children to process complex and abstract information. Because all children are assumed to acquire the same capacities to process rules in about the same order, the appropriate way to test this view involves looking at the average performance of children in different age groups. The average performance is assumed to reflect the underlying mechanism, and individual variations are assumed to be uninteresting.

According to an alternative way of thinking about development, representations have a smaller grain size and only nonenduring states (Thelen, 1995; Thelen & Smith, 1994). On this view, what children learn at some new point depends critically on their previous knowledge. Because children differ in their experience and knowledge, their developmental patterns differ as well. This *dynamic systems* view focuses not on stages of development, but rather on longitudinal data in which the patterns of changes that individual children go through are important. The variability in a single child's performance becomes interesting data to be explained in addition to the stable patterns observed by averaging the performance of many children on a given task. In this way, people adopting symbolic and dynamic systems views of development are led not only to different explanations of behavior but also to different kinds of experiments.

Another case in which assumptions about representations affect the way a psychological process is studied comes from models of the acquisition of past tense forms of English verbs. A key finding to be explained is that early in development, children use both regular and irregular past tense forms properly but then go through a stage of overregularization of the rule, in which they sometimes use the regular past tense form even for irregular verbs (e.g., *goed* instead of *went*). Finally, children learn when to use the regular past tense ending and when not to use it. This pattern of use is reflected in a U-shaped accuracy function in which children start out using irregular past tense forms accurately, get less accurate, and then use the irregular past tense forms properly again. Many models of this process have assumed that past tense learning takes place by learning a set of rules. Because these models are concerned with the formation of rules, the relevant experimental questions are those about the information underlying rule formation, the way that new rules are added, and the way that people learn exceptions to these rules.

An alternative theory was developed by Rumelhart and McClelland (1986), who argued that a representational system sensitive to the statistical structure of the input the child received could explain the observed use of past tense forms. Specifically, they developed a connectionist model of past tense learning (see chap. 2 for more on connectionism). This model also exhibited a pattern of overgeneralization like that observed in children, even though it contained no rules. Rather, the model assumed that children first learned a small number of verbs, most of which were irregular, followed by a larger number of verbs, most of which were regular. At the point at which the large number of regular verbs was added, the model showed some overgeneralization. The learning procedure that the model used was sensitive to these changes in relative frequency of regular and irregular verbs. Subsequent analyses of this model have suggested that it does not capture the fine details of the data (Pinker & Prince, 1988), but the adequacy of this model as an explanation of the development of past tense forms in English is not an issue here. The point is that testing these models experimentally requires examining the statistical structure of the linguistic input that a child receives rather than focusing on the rules that characterize the child's behavior. Thus, the choice of representation embedded in a model (in this case, a distributed connectionist model) leads to a change in the evidence collected. Representational choices affect not only the way cognitive scientists think psychological systems work, but also the way they think about psychology.

This discussion suggests that an important criterion for the adequacy of a proposal about representation is whether it leads to new and interesting experimental questions. I might favor a particular model of mental representation on pragmatic grounds, not on empirical grounds. Ultimately, pragmatic factors of this type have had an important influence on many debates about representation. Eventually, information about how the brain implements cognition will also influence decisions about representations. To think about how the implementation of a process affects the way people think about representation, I first address the general idea of levels of description of cognitive systems.

## LEVELS OF DESCRIPTION

Although I posit mental representations in the process of creating descriptions of psychological processes, there are many ways to describe a system. Both Marr (1982) and Pylyshyn (1980) have explored the idea of levels of description at some length. (In the following discussion, I adopt Marr's terminology, because it is used most often in discussions of representation.) Marr distinguished among *computational*-level descriptions, *algorithmic*-level descriptions, and *implementational*-level descriptions. Each of these describes the same system, but the goal of the description is different in each case.

Imagine a machine that took as input two positive integers and gave as output a third quantity that was the sum of the first two. In what ways can I describe it? At the computational level, a system is characterized simply by the inputs it takes, the output it gives, and the relation between them. In the case of this machine, if I am familiar with arithmetic, I can characterize the machine as taking inputs $a$ and $b$, both of which are positive integers, and providing output $c$ such that $c = a + b$. For many purposes, this is a perfectly good description of the system. I now know the range of allowable inputs to the machine, and I can make accurate predictions of its behavior with those inputs.

Still, there are a number of things I do not know about the system with only a computational-level description. For example, I do not know how the system has chosen to represent these quantities or how it combines these quantities to form the sum. These issues of representation and process are the central aspects of an algorithmic-level description.[5] For example, the machine may use binary notation as a representation. When told that the value of its first input is 8, it first translates the decimal notation given as input into the internal binary representation 100. When told that its second input is 4, it translates this quantity into the internal representation 010. Knowing the way quantities are represented is only half of the battle, though. I must also describe how the represented quantities are combined to form the sum. There may be many ways that the information represented in a particular fashion can be combined to form the sum. The builder of the machine must decide how the representations are combined.

In the case of this machine, it might first contain instructions to look at the right-most digit of each representation. If both digits are 0 or 1, then the right-most digit should be 0; otherwise, it should be 1. If both digits are 1, then a *carry* digit must be set: The machine should look at the next digit to the left in each input quantity, as well as at the carry. The machine may have a table like the one in Table 1.1, which tells it how to set the output digit and carry digit with different patterns of input. For example, with the input $a = 1$, $b = 0$, carry $= 0$ shown in the fifth row

[5] The term *algorithmic level* is a bit misleading from my point of view. Formally, an algorithm is a particular computational procedure frequently associated with digital computers. Specifically, it is the kind of procedure that, when run on a generalized computational device (a Turing machine), is guaranteed to complete its processing in a finite amount of time. The term *algorithm* can also refer to any effective procedure. Only this latter type of algorithm (one that may not halt in a finite amount of time) is central to algorithmic-level descriptions.

TABLE 1.1
Sample Lookup Table for Adding Machine

| *Input* | | | *Output* | |
|---|---|---|---|---|
| *a* | *b* | *Carry* | *c* | *Carry* |
| 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 1 | 1 | 0 |
| 0 | 1 | 0 | 1 | 0 |
| 0 | 1 | 1 | 0 | 1 |
| 1 | 0 | 0 | 1 | 0 |
| 1 | 0 | 1 | 0 | 1 |
| 1 | 1 | 0 | 0 | 1 |
| 1 | 1 | 1 | 1 | 1 |

of the table, the output digit would be set to 1, and the carry digit would be set to 0. The machine then continues by moving to the next input digit to the left in both input quantities until it reaches the left-most digit of the larger quantity.

Choosing this process has many implications. For example, one can predict how long the machine should take to finish combining two numbers. As the numbers to be added get larger, the machine should take longer to add them. In fact, it adds one time step for each new binary digit that it adds to the numbers. Other choices of algorithms make different predictions about timing. For example, the arithmetic and logic units in modern computers add a number of binary digits in parallel. As long as the two inputs do not exceed some value, the addition takes place in a constant time.

For many purposes, the algorithmic level of description is also adequate, but things are still missing from this description. How is the machine constructed? Is it made from silicon chips, pipes filled with water, gears and levers, or animals systematically making marks in the sand? Nothing in either the computational-level description or the algorithmic-level description provides this information. Rather, the physical stuff that makes the system work is the province of implementational-level descriptions. An implementational-level description specifies what the machine is made of. If the machine was constructed from integrated circuit chips, the description would specify the configuration of logic gates. It would state that the binary digit 1 corresponded to a particular voltage in a logic circuit and the binary digit 0 corresponded to another voltage level. It would further specify a mechanism for carrying out the procedure, for allowing the machine to focus its "attention" on the right-most digit and then sequentially to each digit to the left of that one. The implementational-level description is not complete until a working machine could be constructed that would carry out the computation described at the computational and algorithmic levels.

The same computation can be implemented in many ways, although there may be particular time, space, energy, or stability benefits to particular designs. For example, both the old stereo that was in my bedroom when I was growing up and the stereo that is now in my house play vinyl records and produce music. The stereo in my childhood bedroom had vacuum tubes that took a while to warm up. The first album played on any day always sounded a bit worse than the rest of the records, because the system had not reached peak performance. With the advent of solid-state technology, the same function is carried out in silicon chips that do not have to warm up; now even the first record sounds good.[6]

In sum, there are three different levels for describing a mental system. Each level offers a different description of the same system. The computational level specifies the input taken by the system and the output it produces. This input and output need not be symbolic as in addition; they can be a set of forces acting on a limb, with the output being the movement of the limb. The algorithmic level specifies the representations adopted by the system and the processes that extract and use the information from these representations to carry out the function described in the computational-level explanation. The procedure need not be an algorithm that a digital computer can run but simply a process that acts on the representations used by the system. Finally, at the implementational level, the mechanism that is used by the system is actually described.

## Use of Levels of Description in Psychology

In psychology, descriptions of psychological process often move from a computational level to an algorithmic level down to an implementational level. In Kosslyn's (1994) description of the role of mental imagery in the visual system, he began with computational-level descriptions of the components of the system. In fleshing out this description, he presented algorithmic descriptions of the components and, where possible, evidence of how these computations are actually carried out in the brain. Presumably, psychologists would be satisfied with an explanation of the visual and imagery systems if they understood the relation between brain functioning and the representations and processes used.

[6] Of course, in the interim, modes of representation of music have changed to compact discs (CDs), which use different processes to extract information from the representation as well. Indeed, the entire algorithmic level of description has changed, as we have gone from the analog representation embodied in grooves on a record to the digital representation used by compact discs.

The usual order of development is computational-level descriptions followed by lower level descriptions. However, descriptions of algorithms can suggest that a system is too computationally intensive to be feasible. In this case, other potential algorithms that are viable may change the computational-level description of a system. Likewise, evidence about how a process is implemented in the brain may influence both the algorithmic and computational descriptions of the system. The potential of new brain-imaging techniques for the study of psychology is the opportunity to get information about the implementational level of mental processing that can be used to constrain explanations at the computational and algorithmic levels.

The algorithmic level is most central in this book. This level is concerned with descriptions of systems at the level of representations and processes that act over the representations. Thus, when trying to characterize a psychological process and positing potential representations that are used by this process, people think at an algorithmic level.

The boundaries between the levels of description are not always sharp, and the way people think about representations affects psychological descriptions even when working at the computational or implementational level. An example may help make this point clear.

A significant amount of research, particularly in cognitive psychology, adopts the computational level of description. Much of this work aims at developing mathematical models of cognitive processes that characterize the variables that influence a process and uses the mathematical models to specify how these variables affect the process of interest. Nosofsky's (1986) influential generalized context model of classification (see chap. 8), for example, proposed that when learning to place objects into categories, people store representations of the individual category members they come in contact with. When they encounter a new object, they compare it with the specific exemplars stored in memory and classify the new item based on its similarity to the stored exemplars. A description of the entire model is beyond the scope of this chapter, but the model includes parameters for different aspects of categorization, including weights given to attributes that describe the exemplars and weights given to similar and dissimilar values of the attributes.

To make it possible to fit the mathematical model to data, it is assumed that the attributes of the exemplars (e.g., color, size, and shape) can be treated as independent of each other. In particular, it is assumed that each attribute can be compared to a corresponding attribute of other objects: The color of one object can easily be compared to the color of another object without regard to any other attributes of the objects. The model is meant as a description of the kinds of information that are important to the ability to form categories, even though it does not provide specific processes for making comparisons between new objects and previously stored exemplars or for storing old exemplars.

This model does, however, make some assumptions about the nature of mental representations, for instance, that the attributes describing objects are well characterized as being independent features (see chap. 3). This assumption is not only implicit in the mathematical model but is also carried through to the materials used in empirical studies of categorization. These studies often use simple materials with a small number of easily separable and conceptually independent dimensions. Thus, even though this work is ostensibly directed at forming a computational-level description, it also influences the representations deemed most appropriate for understanding the behavior.

A computational-level description need not place strong constraints on the representations posited at the algorithmic level. Often, many different representations and processes are consistent with a computational-level description (e.g., Barsalou, 1990). In practice, however, the computational-level description does seem to place constraints on how people think about the representations relevant to the process and in turn about what evidence is collected to test the descriptions of the process.

This link among levels of description, representation, and collecting and interpreting psychological evidence is the main motivation for this book. Every decision about how to describe a psychological process carries with it some assumptions about mental representation. Ideas about representation and processing carry with them suggestions about how to collect evidence for this process, and it is important to think carefully about the representations assumed by a description of a psychological process. It is crucial to be aware of making a set of representational assumptions. Better still is making choices about representations because of the particular things that a representation makes easy or hard to do.

## ORGANIZATION OF THIS BOOK

What is the best way to think about mental representation? Researchers have extensively studied the types of representations related to particular tasks. Debates about how people carry out logical reasoning tasks have focused on assumptions about whether people's representations are more sentence-like (e.g., Rips, 1994) or more model-like (Johnson-Laird, 1983; Johnson-Laird & Byrne, 1991). Debates over people's ability to use mental imagery have focused on whether people's visual perception and imagery ability use analog or propositional representations (Cooper, 1975; Kosslyn, 1994; Pylyshyn, 1981). It is tempting to look at particular tasks and to examine the representations used to represent the content in these tasks. I have decided, however, to take a different tack.

Very similar sets of representations have been proposed across many areas of cognitive science. Some of these similarities are well-known, as in the assumption that sentence-like representations are useful for understanding both logical reasoning and mental imagery ability. Other similarities may be less obvious, as those between distributed connectionist models and multidimensional semantic space models of similarity. In an effort to make some of these similarities more apparent, I have organized this book around types of representations.

The book begins with simple representations and moves to representations of greater complexity. To illustrate the uses of many of the models, I describe implemented computer programs in which the representational assumptions have been made explicit. The link between serious thought about representation and computational modeling is important: Computational work can confirm specific predictions of a theory. Indeed, constructing a working model of a process can provide an existence proof that the theory can actually account for a set of phenomena (Kosslyn, 1994). Furthermore, when developing a computational model, both the representation and the process must be fully spelled out. This specificity may enable researchers to make new predictions about the nature of the cognitive process under study.

Chapters 2 to 4 deal with unstructured representations, which contain no explicit bindings or ownership between elements in the representations. Chapter 2 examines spatial representations, chapter 3 featural representations, and chapter 4 associative networks. These representations are simple and have the advantage of carrying with them simple processes that act over them. Despite this apparent simplicity, they are very powerful, and many aspects of cognitive processing may be well modeled using these representations.

Chapters 5 to 7 focus on structured representations. Chapter 5 introduces the notion of binding and presents some general structured representations. Chapter 6 examines how structured representations may be used in perception and how this use may extend to conceptual thought. Chapter 7 looks at scripts and schemas, which are perhaps the most common structured representation incorporated into psychological models.

Chapters 8 and 9 deal with general issues. Many examples in chapters 2 through 7 involve abstractions, like a typical bird or a typical visit to the doctor. Chapter 8 examines how people incorporate information about specific episodes (e.g., a particular doctor's office visit), or specific items (e.g., a particular doctor) into representations. Chapter 9 looks at mental models and tries to clarify the way the term *mental model* has been used in psychology. Finally, in chapter 10, I return to some general issues addressed in chapter 1 and discuss them in light of the representations presented in the book.