CHAPTER 41

The Mental Models Perspective

Philip N. Johnson-Laird

Abstract

This chapter begins with mental models as the end product of vision, as a repository of knowledge, and as underlying visual imagery. It contrasts them with the alternative hypothesis that mental representations are syntactic expressions in a mental language. To resolve this controversy, it shows that the structure of models, which differs from that of syntactic representations, plays a major role in accounting for the comprehension and memory of discourse. It reports evidence corroborating another major principle of mental models—that they normally represent only what is true—for models of concepts and models of propositions. The chapter then describes how intuitions are based on a single mental model, whereas deductions call for the representation of alternative models, especially those representing counterexamples to putative conclusions. It reports the corroboration of these predictions. Next, it turns to inductive reasoning. It shows how models underlie common forms of induction in daily life, and it reports evidence corroborating the prediction that individuals prefer explanations that resolve inconsistencies over minimal amendments to the offending propositions. Finally, it concludes with an overview of the main principles governing mental models.

Key Words: concepts, deduction, explanation, induction, logic, mental models

The immediate precursor to the modern theory of mental models is a hypothesis due to the prescient psychologist and physiologist, Kenneth Craik (1943). What he wrote conveys the essence of the modern theory:

> If the organism carries a "small-scale model" of external reality and of its own possible actions within its head, it is able to try out various alternatives, conclude which is the best of them, react to future situations before they arise, utilize the knowledge of past events in dealing with the present and the future, and in every way to react in a much fuller, safer, and more competent manner to the emergencies which face it. (Craik, 1943, p. 61)

Before Craik, various philosophers and physicists had proposed analogous ideas. The great 19th-century American logician C. S. Peirce, for example, postulated that reasoning depends on diagrams that are models of propositions (Peirce, 1931–58, Vol. 4). Readers can find a fuller history of mental models elsewhere (Johnson-Laird, 2006), and so the aim of the present chapter is to describe the modern theory, which began to be formulated about 30 years ago, and which has had an impact on various aspects of cognitive psychology—from the study of perception to high-level reasoning.

This chapter begins with mental models as the end product of vision, as a repository of knowledge, and as underlying visual imagery. It contrasts them with the alternative hypothesis that mental representations are syntactic expressions in a mental language. To resolve this controversy, it shows that the structure of models, which differs from that of syntactic representations, plays a major role in accounting for the comprehension and memory of discourse. It reports evidence corroborating another major principle of mental models—that they normally represent only what is true—both for models of concepts and for models of propositions. The chapter then describes how intuitions are based on a single mental model, whereas deliberations call for the representation of alternative models, especially those representing counterexamples to putative conclusions. It reports the corroboration of these predictions. Next, it turns to inductive reasoning. It shows how models underlie common forms of induction in daily life, and it reports evidence corroborating the prediction that individuals prefer explanations that resolve inconsistencies over minimal amendments to the offending propositions. Finally, it concludes with an overview of the main principles distinguishing mental models from other sorts of putative mental representation.

The Modern Theory of Mental Models

You may have the intuition that vision, hearing, and your other senses put you into direct contact with reality. In fact, you have no such contact. Consider, for example, the old riddle: If a tree falls down in the middle of a forest, miles from any sentient entity, does it make a sound? The answer is: no. It makes vibrations in the air, but sound itself depends on hearing. Similarly, objects in the world reflect light, but colors, textures, and shapes depend on a visual system. Vision, as Marr (1982) argued, is an unconscious inference, starting from the patterns of light falling on your retinas, and leading to a mental model that makes explicit the three-dimensional structure of the scene in front of you. To move around safely, you need a representation of the world that is independent from your viewpoint: You need to know what things are where in the world. You can recognize, say, that a street contains shops, pedestrians, and passing cars; and you can readily make your way to a distant landmark even if you have never been on that particular street before. Vision solves three problems to enable you to do so: It constructs a mental model that makes explicit three-dimensional shapes, it uses these shapes to identify objects, and it makes explicit the spatial relations among them.

How these tasks are carried out is not known with any certainty. Marr supposed that you have a catalog of the three-dimensional mental models of familiar objects, and that your visual system computes the shape of entities in the scene in terms of their major axes, for example, a furled umbrella is a long tapering cylinder. It compares this shape with the shapes in its catalog, which at their highest level capture the overall shape of objects but at lower levels flesh out the detailed shapes of the parts of objects. One possibility is that a visual cue about the shape of an object may trigger access to a model in the catalogue, which is then used to try to match the rest of the percept (cf. Biederman, 1995; Ullman, 1996; for alternative hypotheses).

The term "mental model" is often used in a promiscuous way to refer to any systematic representation of knowledge (Gentner & Stevens, 1983). Such a model of the world could have a three-dimensional structure (Hegarty, 1992; Metzler & Shepard, 1982), or it could consist of propositional representations, which are syntactically structured expressions in a mental language (Pylyshyn, 2003). This latter view dovetails with theories of reasoning in which formal rules of inference akin to those of logic are applied to representations of the logical form of propositions (e.g., Braine & O'Brien, 1998; Rips, 1994). In stark contrast, other psychologists have argued for theories that eschew abstract representations in favor of ones rooted in perception (Barsalou, 1999; Markman & Dietrich, 2000). Surprisingly, the resolution of this controversy comes, not from the study of perception or imagery, but from investigations of language and reasoning. They offer a more precise notion of mental models, and they provide evidence corroborating their psychological reality.

Mental Models and Comprehension

The Interpretation of Discourse

You construct models of the world from perception, but you can also construct them from descriptions of the world, which enable you to experience it by proxy. A good writer or storyteller has the power to initiate a process similar to the one that occurs when you perceive or imagine events. Indeed, experiments show that individuals rapidly forget the surface form of sentences, their underlying grammatical relations, and even the gist of individual sentences. And so when you understand discourse, you use the meaning of sentences and your general knowledge to construct mental models of the situations under description (e.g., Johnson-Laird, 1983; Van Dijk & Kintsch, 1983). But what distinguishes a mental model from, say, a propositional representation in a mental language?

The answer according to the present author is that a mental model has a structure corresponding to the structure of what it represents (Johnson-Laird, 1983). It is *iconic.* That is, its parts are interrelated in the same way that the parts of the entities that it represents are interrelated (see Peirce, 1931–1958, Vol. 4, paragraph 433 for this notion of iconicity). Models accordingly represent what things are where in a visual scene or in its verbal depiction, though in the latter case the model is compatible with an indefinite number of scenes. The model represents each referent with a single mental token, the properties of referents with properties of the tokens, and the relations among referents with relations among the tokens. This property of iconicity therefore distinguishes mental models from other sorts of representation, such as those in a mental language, which have a syntactic structure rather than an iconic one.

To illustrate the iconicity of a mental model, consider a simple spatial description (see Byrne & Johnson-Laird, 1989):

The talk button is on the left of the close-doors button. The open-doors button is
on the right of the close-doors button.

Your interpretative system constructs a representation of the meaning of each sentence, and it can use this meaning to construct or to update a mental model of the spatial layout of the buttons. This model is depicted in the following diagram in which the left-to-right axis corresponds to that of the panel of buttons:

Talk Close-doors Open-doors

As the diagram illustrates, the model is iconic in that its layout corresponds to the layout of the three buttons, but a mental model represents actual buttons on an elevator, not just their verbal labels. You could use the model to infer that the talk button is to the left of the open-doors button. No alternative model of the description is a counterexample to this conclusion, and so it must be true given the truth of the description.

Experimental evidence shows that the number of mental models that individuals need to construct to make an inference predicts its difficulty, whereas the length of a logical proof based on propositional representations does not (e.g., Byrne & Johnson-Laird, 1989, Johnson-Laird & Byrne, 1991). Likewise, other evidence suggests that mental models underlie memory for descriptions (Bransford, Barclay, & Franks, 1972; Garnham, 1987). Mental models of a story can be dynamic and unfold in time (Johnson-Laird, 1983, Ch. 6), and Oatley and his colleagues have argued that fiction is a device for creating such simulations (e.g., Mar & Oatley, 2008). As a corollary, changes in location in a story should affect your ease of accessing the various individuals and entities in the story. For example, if the protagonist walks from one room to another carrying an object, then it is easier for you to access this object and the entities in the new room than those in the room the protagonist has just left. It takes you longer to respond to questions or to a probe word about them; and similar effects occur for stories (e.g., Glenberg, Mayer, & Lindem, 1987; Rinck & Bower, 1995), movies (e.g., Magliano, Miller, & Zwaan, 2001), and "virtual reality" (Radvansky & Copeland, 2006). Hence, you maintain a model of discourse similar to one that you construct from perceiving the events.

Spatial relations, such as those in the earlier description of the buttons in an elevator, are easy to envisage. You might therefore assume that mental models are nothing more than visual images. This assumption is wrong. Some descriptions are easy to visualize yet do not elicit spatial representations, for example, "The dog is dirtier than the cat." When individuals reason from such propositions, they are slower in comparison with their reasoning from propositions that do not elicit images (Knauff & Johnson-Laird, 2002). The reason may be that only descriptions eliciting imagery activate regions in visual cortex (Knauff, Fangmeier, Ruff, & Johnson-Laird, 2003). Some propositions, such as "Ann is cleverer than Beth," are plainly impossible to represent solely in a visual image. You can imagine, say, Ann as higher on a vertical scale than Beth, but nothing in such a representation makes explicit the meaning of *cleverer than*. Not all properties or relations are rooted in a sensory modality. Mental models can contain abstract elements, which, as Peirce realized, are symbolic rather than iconic.

The idea that discourse is represented in mental models of the situations under description is relatively uncontroversial (see, e.g., Gernsbacher, 1990; Kintsch, 1988). The major problem for the system that builds such models is to establish the appropriate referent for each expression. Speakers refer back to entities that they have already introduced in the discourse, and they can use different noun phrases, demonstratives, or pronouns to do so. The interpretative system uses many cues to co-reference—from the meaning of sentences to the grammatical roles of noun phrases (Almor, 1999; Stevenson, Nelson, & Stenning, 1995). The most comprehensive account within the framework of mental models is due to Garnham and his colleagues (e.g., Cowles & Garnham 2005; Garnham, 2001). It postulates that a critical factor is the number of potential antecedents for a referring expression, and so a noun phrase needs enough content to pinpoint its antecedent among them. But a noun phrase can also signal the future direction of the discourse and perhaps a shift in theme. No current theory, however, has led to a theory comprehensive enough to yield a computer program that copes with natural language.

A mental model captures what is common to the different ways in which a possibility can occur, and so the theory is an analog of "possible world" semantics (Kripke, 1963) and its more recent variants such as "situation semantics" (Barwise, 1987) and "discourse representation" theory (Kamp & Reyle, 1993). But representations according to these theories are always correct, whereas, as the next section shows, mental models have intrinsic shortcomings that lead individuals into error.

The Principle of Truth

A central assumption of the model theory is known as the principle of *truth*. It postulates that mental models represent only what is true (Johnson-Laird & Savary, 1999). The principle is subtle, because it operates on two levels. Given an "exclusive" disjunction, such as:

> Either the man pressed the open-doors button or else the woman pressed the close-doors button, but not both

a truth table lays out the truth or falsity of the disjunction for all four possible contingencies depending on the truth or falsity of its two constituent clauses. The disjunction is true when one clause is true and the other clause is false, and otherwise it is false. In contrast, the mental models of the proposition represent only the two contingencies that are possible given the truth of the proposition, which are laid out in this diagram on separate lines:

man pressed open-doors button
woman pressed close-doors button

For convenience, sentences stand in for mental models of actions in these diagrams. At the second level of the principle of truth, the models above do not represent explicitly that in the first possibility it is false that the woman pressed the close-doors button, and that in the second possibility it is false that the man pressed the open-doors button. The theory allows, however, that certain circumstances may lead individuals to flesh out their mental models into *fully explicit* models, which use negation to represent the status of each clause in each possibility:

man pressed open-doors button & woman did *not* press close-doors button
man did *not* press open-doors button & woman pressed close-doors button

The principle of truth reduces the load on working memory in comparison with a truth table, which represents all four possible contingencies. And the principle seems benign. Yet it can lead individuals into the illusion that they understand a description when, in fact, they have misunderstood it. A computer program implementing the principle led to the discovery of a variety of such illusions, and subsequent studies have corroborated their occurrence (Johnson-Laird, 2006). The problem arises when mental models of one proposition fail to take into account the concurrent falsity of another proposition, and so the mental models differ from the correct fully explicit models of the description.

A striking illusion of this sort occurs with the description:

> Suppose only one of the following assertions is true:
>
> 1. You have the bread.
> 2. You have the soup or the salad, but not both.
> Also, suppose you have the bread. What, if anything, follows? Could you have both the soup and the salad?

Most participants (78%) say, "no" (Khemlani & Johnson-Laird, 2009), and this answer is predicted by the mental models of the two initial assertions, which yield three alternative possibilities for what you can have:

bread
soup
salad

In contrast, the fully explicit models, which represent the status of each clause in the description in each model, are quite different. Given that you have the bread, then assertion 1 is true, and so assertion 2 is false. And there are two ways in which it can be false. One way is that you have neither the soup nor the salad, but the other way is that you have both of them. The fully explicit models of the description are accordingly:

bread & no soup & no salad
bread & soup & salad

where each model is a conjunction of entities. So, the correct answer to the question is: "Yes, given that I have the bread, I can also have both the soup and salad."

MODELS OF CONCEPTS

An earlier part of the chapter described how models of entities can be part of the models of a visual scene. Similarly, models of concepts are components of the models of propositions, for example, your model of the concept of, say, "soup" is part of your model of the assertion, "You have the soup." You begin to acquire concepts in infancy and continue to do so, and to devise novel concepts, throughout your life (Medin, Lynch, & Solomon, 2000). One way to create new concepts is by combining existing ones using logical connectives, such as negation, conjunction, and disjunction. For example, the concept of a "ball" in baseball is defined as a pitch at which the batter does *not* swing *and* which does *not* pass through the strike zone. Systems based on these connectives, and those that can be defined in terms of them, are known as "Boolean" in honor of George Boole, the logician who first systematized their logic. But even informal concepts often depend on Boolean connectives, for example, the relation of ownership, as in *she owns it*, means in part that it is permissible for her to use it, *and* it is *not* permissible for others to prevent her from using it (Miller & Johnson-Laird, 1976, p. 560).

If thought depends on representations in a mental language, Boolean concepts should depend on them too, with expressions of the form, for example, *a and b*, *or not-a or not-b*. Likewise, the acquisition of a concept should call for individuals to find a minimal Boolean description of the instances of a concept (Feldman, 2000), or some sort of logical description of them (Nosofsky, Palmeri, & McKinley, 1994; Vigo, 2009). The resulting description could yield a decision tree that yields a correct classification of instances and noninstances of the concept (e.g., Bruner, Goodnow, & Austin, 1956; Hunt, 1962; Shepard, Hovland, & Jenkins, 1961). In contrast, the model theory postulates instead that a concept is represented in mental models of its possible instances, which are each a conjunction of properties and relations (Goodwin & Johnson-Laird, 2012). For example, the concept *tall and thin, or else short and fat* has these two mental models of its instances:

tall   thin
short   fat

Each model represents one sort of possible instance consisting of a conjunction of attributes, but here and henceforth, for simplicity, the sign for conjunction, "&", is omitted from these diagrams.

The simplest way to acquire a concept is to commit to memory each of its exemplars (see, e.g., Medin & Smith, 1984). The model theory, however, postulates that individuals detect those attributes that are irrelevant given the values of other attributes, and they then eliminate the irrelevancies (Goodwin & Johnson-Laird, 2010). As an example, consider a concept that has these two instances:

tall thin muscular
tall thin not-muscular

Clearly, the attribute of *muscular or not* is irrelevant to the concept, which can be represented in a single model:

tall thin

The particular simplifications that humans discover are likely to depend on the order in which they encounter the instances of a concept, and on the relative saliency of their attributes. However, the overall number of models that result from the elimination of irrelevant attributes does not change as a result of these differences, and it provides a better predictor of the difficulty of acquiring concepts than either the number of decisions in a decision tree (Hunt, 1962) or the length of a minimal description of concepts (Feldman, 2000). So when individuals learn to categorize instances and noninstances of concepts, they do not seek a minimal description of the concept but instead seek to minimize the number of mental models required to represent its instances (Goodwin & Johnson-Laird, 2012). They eliminate any irrelevant property or relation. They also base their *descriptions* of a concept on mental models of its instances. That is, they describe disjunctions of instances, omitting irrelevant attributes.

If concepts are represented in models, then illusory concepts should exist, and recent studies have corroborated their existence (Goodwin & Johnson-Laird, 2010). Consider, for instance, this description of a set of objects based on their color and shape:

red if and only if square, or else red.

The description yields two mental models:

red   square
red

Hence, individuals think that the concept includes red squares. But the fully explicit models of the concept show that this concept is illusory:

not-red not-square (the first clause of the disjunction holds, but the second does not)

red not-square (the first clause does not hold, but the second does)

Readers may think that such concepts are highly artificial, and that errors are merely a consequence of this artificiality. A simple control inference, however, is just as artificial. It depends on changing the disjunction to an *inclusive* one, which allows that both its clauses could be true. In this case, the mental models yield the correct answer, and individuals tended to make it, too. Performance was also good on other control inferences based on exclusive disjunctions. The following description:

> red and green, or else green.

should not elicit the illusory model:

> red green

because individuals know that the objects under description cannot be both red and green. An experiment corroborated this prediction. Individuals were much less likely to succumb to illusions when the content of the descriptions blocked an illusory model, leaving only a correct model of the concept (Goodwin & Johnson-Laird, 2010, Experiment 3). Content had only a small effect on performance when it blocked an illusory model, but the participant still had to recover the correct models. And it had no effect whatsoever when it blocked one illusory model but not another. No other current theory, including recent probabilistic accounts (Kemp & Tenenbaum, 2008), predicts the occurrence of illusory concepts. Hence, their occurrence is a crucial corroboration of the model theory.

Logical Reasoning

Deduction and Logic

Reasoning is a systematic mental process that generates or evaluates implications among propositions. Implications are of two main sorts: deductive and inductive. Deduction is a central cognitive process and a major component of intelligence (Stanovich, 1999), and so tests of intelligence include problems of deductive reasoning. You know, for instance:

> If one earns a salary, then one pays income tax.
> President Obama earns a salary.

And so you can infer:

> President Obama pays income tax.

This inference is *valid*, that is, if its premises are true, then its conclusion must be true, too. Logicians define a valid inference as one whose conclusion is true in every possibility in which all its premises are true (Jeffrey, 1981, p.1). In other words, there are no counterexamples to a valid deduction, that is, no possibilities in which the premises hold but the conclusion does not.

Psychologists studying reasoning once aimed to identify the particular logic that people have in their heads—an idea going back to the ancient doctrine that the laws of logic are the laws of thought. One difficulty was the vast number of different logics, including the indefinitely many "modal" logics for possibility and necessity. Nevertheless, theorists argued for a century that logic is a theory of human deductive competence; and Inhelder and Piaget (1958, p. 305) proposed that reasoning is nothing more than logic itself. Others have similarly argued that deductive performance depends on formal rules of inference (e.g., Braine & O'Brien, 1998; Rips, 1994). But several difficulties confront any psychological theory based on logic. Some are theoretical, such as the fact that in logic infinitely many conclusions—most of which are trivial—follow validly from any set of premises, whereas individuals often say, quite sensibly, that nothing follows from certain premises. Logic has nothing to say about which logical conclusions are sensible. What naïve individuals—those who have not mastered logic—tend to infer are conclusions that do not add disjunctive alternatives to those possibilities to which the premises refer, that simplify matters rather than include redundant propositions, and that make explicit what was only implicit in the premises (Johnson-Laird & Byrne, 1991, p. 22).

Another problem for theories based on logic is the difficulty of establishing the logical *form* of everyday propositions. In logic, deductions are expressed in sentences in a formal language with a grammar that makes logical form explicit, and they are proved using rules of inference sensitive only to these logical forms. But, in everyday life, implications hold, not between sentences, but between the propositions that sentences express in a particular context, or propositions that derive from perception, memory, or imagination. What proposition an everyday sentence expresses depends on its meaning, on what it refers to, and on knowledge. The one computer program implementing a psychological theory based on formal rules accordingly calls for users themselves to provide the logical form of the premises and conclusion (Rips, 1994). So what is an alternative basis for reasoning?

Craik postulated that models help us to navigate our way through life, but he did not consider their role in reasoning, which he took to depend on "verbal rules"—an idea on which he did not elaborate (Craik, 1943, p. 81). Models, however, are a way in which to make inferences. Reasoners construct models based on descriptions, on perception, and on knowledge. They formulate a conclusion that holds in the models and that was not overtly asserted in any single premise. A conclusion that holds in all the models is necessary given the premises (Johnson-Laird & Byrne, 1991). A conclusion that holds in most of the models is probable given the premises (Johnson-Laird, Legrenzi, Girotto, Legrenzi, & Caverni, 1999). And a conclusion that holds in at least one model is possible given the premises (Bell & Johnson-Laird, 1998). Models accordingly provide a unified theory of logical, modal, and probabilistic reasoning—at least the sort of probabilistic reasoning that depends on adding the probabilities of the different ways in which an event can occur. In fact, models have been successfully applied to most aspects of everyday deduction (for a review, see Johnson-Laird, 2006), but here the focus is on sentential reasoning, which depends on negation and connectives, such as "if," "or," and "and," that is, the same connectives that are used in describing Boolean concepts (see the earlier section).

Intuitions and Deliberations: A Dual-Process Account

The modern theory of mental models from its inception differentiated between intuitions and deliberations (Johnson-Laird, 1983, Ch. 6). On the one hand, you make rapid, effortless, and unconscious inferences. For example, you read the following description:

> There was a fault in the signaling circuit. The crash led to the deaths of two people.

You infer that the crash killed them. The text makes no such assertion, and it could continue:

> They were arrested after the accident, convicted of deliberately causing the fault, and shot as saboteurs.

On the other hand, you make voluntary, effortful, and conscious inferences that take time. Psychologists have largely focused on these inferences at the expense of implicit inferences, which were discovered by computer scientists trying to write programs that "understand" natural language. The crucial difference between the two sorts of inference according to the model theory is that intuitive inferences depend on "a *single* mental model [based on] the discourse, its context, and background knowledge" (Johnson-Laird, 1983, p. 128). No attempt is made to search for alternative models unless evidence occurs to overrule the model. Hence, the process can be rapid and unconscious, but there is no guarantee that its results are valid. In contrast, deliberate reasoning depends on working memory and on carrying out recursive processes, including a search for alternative models. Hence, the process is slow and you are aware of reasoning.

Many psychologists have proposed such "dual-process" accounts of reasoning (e.g., Johnson-Laird & Wason, 1976, p. 5–6; Kahneman & Frederick, 2002; Rader & Sloutsky, 2002; Schroyens, Schaeken, & Handley, 2003; Sloman, 1996; Stanovich, 1999). But Evans and his colleagues have perhaps explored the idea in more depth than other investigators (e.g., Evans, 2003; and Evans & Over, 1996; Wason & Evans; 1975). Not all theories specify how the two sorts of reasoning work together or what the processes are on which they rely. Such an algorithm, however, is built into those programs implementing the model theory (e.g., Johnson-Laird & Byrne, 1991, Ch. 9): the intuitive process constructs a single model of the premises, and the deliberative process searches recursively for alternative models.

A conditional assertion such as:

If one earns a salary, then one pays income tax

usually refers to three possibilities:

earns salary	pays income tax
doesn't earn salary	pays income tax
doesn't earn salary	doesn't pay income tax

The Queen of England is an example of the second possibility. When you understand such a conditional, you normally construct only one explicit mental model that represents the most salient possibility—the first one in the list above, and another model with no explicit content to allow for the other possibilities. The further assertion, say, that Obama earns a salary, eliminates this implicit model, leaving only the explicit mental model, and it suffices for you to infer that Obama pays income tax. In the different case in which, say, that Charles does *not* pay income tax, your mental models of the conditional yield no conclusion, and a common error is to think that nothing follows from such premises. When you deliberate, however, you can flesh out your mental models into fully explicit models representing all three possibilities above. Now, you can infer from the premise about Charles that he does not earn a salary (see Verschueren, Schaeken, & d'Ydewalle, 2005). Oberauer (2006) showed that this "dual-process" theory of mental models gives a better account of reasoning from conditional assertions than its rivals.

Models and Sentential Reasoning

The theory of mental models yields five main predictions about sentential reasoning. First, more models mean more work; that is, the greater the number of models of possibilities that you need to think about, the harder an inference will be. Second, you can use counterexamples to overturn invalid inferences. Third, the principle of truth, which was described earlier, implies that you should make illusory inferences. Fourth, you can develop various strategies for reasoning, but, regardless of your strategy, the previous predictions should still hold. And, fifth, the meaning of clauses and general knowledge can modulate your interpretation of sentential connectives, such as "if" and "or," so that they no longer refer to three possibilities illustrated earlier. This section of the chapter examines each of these predictions in turn.

MORE MODELS MEAN MORE WORK

The greater the number of models that individuals have to think about, the harder deductions should be, taking longer and being more prone to error. These errors should consist in drawing conclusions that overlook at least one model of a possibility consistent with the premises. A corroboration of this prediction concerns the difference in reasoning from an exclusive disjunction, for example:

> The man pressed a button or else the woman pressed a button, but not both.

and in rreasoning from an inclusive disjunction, for example:

> The man pressed a button or the woman pressed a button, or both.

As we saw earlier, an exclusive disjunction refers to two possibilities, whereas an inclusive disjunction refers to three possibilities, because it allows that both of its clauses could be the case. Granted, say, the further premise:

> The man did not press a button

it is a valid inference from either sort of disjunction that:

> The woman pressed a button.

The model theory predicts that the inference from the exclusive disjunction (two possibilities) should be easier than the inference from the inclusive disjunction (three possibilities). The theories based on formal logic make the opposite prediction, because they have a rule for inclusive disjunction, but not for exclusive disjunction, and so the inference from it calls for a sequence of steps. Reasoning from verbal premises and from diagrams has corroborated the model theory's prediction (Bauer & Johnson-Laird, 1993; Johnson-Laird, Byrne, & Schaeken, 1992). Likewise, reasoning from a conjunction, which refers to only one possibility, is easier than reasoning from a disjunction (García-Madruga, Moreno, Carriedo, Gutiérrez, & Johnson-Laird, 2001).

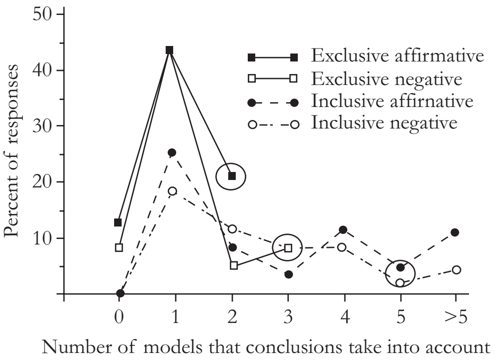

Figure 41.1 The number of models of the premises underlying the participants' conclusions from four sorts of pairs of disjunctive premises (from Johnson-Laird et al., 1992). The circled items are the correct valid conclusions.

The erroneous conclusions that individuals draw tend to hold for only some of the possibilities to which the premises refer. Figure 41.1 presents the number of models of the premises that the participants' conclusions took into account in disjunctive reasoning (Johnson-Laird et al., 1992). The premises were pairs of either inclusive or exclusive disjunctions, and the second disjunction contained a clause that either affirmed or negated a clause in the first disjunction. The circled points in the figure correspond to the valid conclusions, which depend, respectively, on two, three, or five models. The participants drew just over 20% of valid conclusions for the two-model inferences, and less than 10% for the remaining inferences. And, as the figure shows, the modal errors were conclusions consistent with only one model, that is, the participants overlooked many possibilities to which the premises referred. Formal rule theories make no use of representations of possibilities, and so they cannot explain these results.

COUNTEREXAMPLES OVERTURN INVALID INFERENCES

A counterexample to an inference is a possibility in which the premises hold, but the conclusion does not. There are two main sorts of invalid inference: one sort yields a conclusion that is consistent with the premises but that does not follow from them, for example:

> The fault is in the cable or the printer, or both.
> Therefore, the fault is in the cable and the printer.

The conclusion holds in one possibility to which the premise refers, but not in all of them, and so it is invalid, as the following counterexample shows:

> fault in cable & not a fault in the printer.

The other sort of invalid inference yields a conclusion that is not even consistent with the premises—it is impossible given that they hold, for example:

> The fault is in the cable or the printer, or both.
>
> Therefore, the fault is not in the cable and not in the printer.

The best way in which to elicit counterexamples from naïve individuals is to ask them to evaluate given conclusions that are invalid but consistent with the premises. When the participants write justifications for their correct evaluations of such inferences, they tend to describe counterexamples (Johnson-Laird & Hasson, 2003). Studies of other sorts of reasoning have also shown that individuals use counterexamples spontaneously in drawing their own conclusions (e.g., Bucciarelli & Johnson-Laird, 1999). And, as brain imaging showed, only a search for counterexamples triggered activity in the region of the right frontal hemisphere known as the "frontal pole" (Kroger, Nystrom, Cohen, & Johnson-Laird, 2008). Psychological theories based on formal rules make no use of counterexamples, and so cannot account for these results.

ILLUSORY SENTENTIAL INFERENCES

Readers are invited to solve the following problem:

> Either Jane is kneeling by the fire and she is looking at the TV or otherwise Mark is standing by the window and he is peering into the garden.
> Jane is kneeling by the fire.
> Does it follow that she is looking at the TV?

Most people say, "yes" (Walsh & Johnson-Laird, 2004). The computer program implementing the model theory predicts that the premise yields two mental models:

> Jane: kneeling by fire and looking at TV
> Mark: standing by window and peering into garden

Hence, the theory predicts the affirmative answer. In fact, the answer is wrong. If the second conjunction is true, then the first conjunction is false, and one way in which it could be false is that Jane is kneeling by the fire but *not* looking at TV. This example is a paradigm case of an illusory inference, and the present author should confess that he succumbed to it when Walsh and he were designing the materials for an experiment. A common criticism of such studies is that the materials are artificial, and so why should we care about their results? The principal answer is that reasoning is often about artificial contents both in logic and in daily life—the worldwide popularity of Sudoku puzzles is an excellent example (Lee, Goodwin, & Johnson-Laird, 2008). Their solution depends on pure deductive reasoning, but their contents are utterly artificial. In studies of illusory inferences, we can be sure that neither the contents nor the framing of problems causes such poor performance, because the participants are highly accurate in responding to the control problems.

Illusory inferences occur with all sorts of sentential connective, including "or" in both its inclusive and exclusive senses, and "if" and "if and only if" (e.g., Johnson-Laird & Savary, 1999). They also occur in various other domains of reasoning (e.g., Bucciarelli & Johnson-Laird, 2005; Goldvarg & Johnson-Laird; 2000; for a review, see Johnson-Laird, 2006). Many experts have fallen for them, and then proposed ingenious explanations for their errors, for example, the premises are so complex or artificial that they confuse people. But reasoners are highly confident in their conclusions, and, as I mentioned earlier, the control inferences, which participants get right, are equally complex and artificial. Other putative explanations concern the interpretation of conditionals. But the illusions occur with disjunctions too, and their interpretation is not controversial. Certain procedures do alleviate the illusions (e.g., Barrouillet & Lecas, 2000; Santamaria & Johnson-Laird, 2000; Yang & Johnson-Laird, 2000), but a perfect antidote for them has yet to be discovered.

INDIVIDUAL STRATEGIES IN SENTENTIAL REASONING

Readers might suppose that individuals are equipped with a single deterministic strategy for deduction, which unwinds like an algorithm for long multiplication. One reason for this view is that many experiments are insensitive to the use of different strategies. Yet there are long-standing impediments to the notion of a single reasoning strategy. For example, the order of premises has robust effects on inferences—in a way that the model theory predicts (Girotto, Mazzocco, & Tasso, 1997). And when individuals carry out a series of sentential inferences, they develop different strategies for coping with them. This phenomenon is obvious when they think aloud and are permitted to use pencil and paper (Van der Henst, Yang, & Johnson-Laird, 2002). Consider, for instance, the following sort of inference about the contents of a box:

> There is a red marble in the box if and only if there is a brown marble.
> Either there is a brown marble or else there is a gray marble, but not both.
> There is a gray marble if and only if there is a black marble.
> Does it follow that: If there is not a red marble then there is a black marble?

The inference is easy, and the correct answer is "yes." Over the course of several problems of a similar sort, different individuals develop different strategies for reasoning about them.

Some people spontaneously develop a strategy based on suppositions. When they think aloud, they say, for instance:

> Suppose there isn't a red marble. It follows from the first premise (above) that there isn't a brown one. It then follows from the second premise that there's a gray marble. The third premise then implies that there's a black one. So, yes, the conclusion does follow.

Each of these inferential steps can be carried out using models. The participants do not always use suppositions correctly. Given the conclusion in the previous example, participants sometimes made the supposition: suppose there's a black marble. They then inferred from the premises that there is not a red marble, and so they responded that the conditional followed from the premises. They made the correct response, but not for the right reason. The "then" clause of conditional can be true even when its "if" clause is false, and so the right way to proceed is to make a supposition of the "if" clause and to show that it leads to the truth of the "then" clause.

Another strategy is to make an inference from a pair of premises, and then to make another from its conclusion and the third premise. One strategy was totally unexpected, and no previous mention of it appears to be in either the psychological or logical literature. Reasoners transform each premise, where necessary, into a conditional, so that the result is a chain of conditionals leading from one clause in the conditional conclusion to its other clause, for example:

> If there isn't a red marble then there isn't a brown marble.
> If there isn't a brown marble then there is a gray marble.
> If there is a gray marble then there is a black marble.
> So, if there isn't a red marble then there is a black marble.

The strategy is correct provided that reasoners make the correct transformations into conditionals, and that they construct a chain leading from the "if" clause of the conclusion to its "then" clause. However, they sometimes worked incorrectly in the opposite direction. The model theory predicts that it should be easier to make inferences from conditionals than from disjunctions (see also Ormerod & Richardson, 2003), because conditionals have only one explicit mental model, whereas disjunctions have at least two explicit mental models. Hence, the construction of chains of conditionals should be much more likely than the construction of chains of disjunctions. Indeed, not a single participant ever transformed a conditional into a disjunction.

The most frequent strategy was to draw a single diagram that represented all the premises. For example, some participants drew a horizontal line across the middle of the page and wrote down the two possibilities to which the premises referred:

red brown
___
gray black

A tell-tale sign of this strategy is that individuals work through the premises in whatever order they are stated, even taking into account irrelevant premises. When individuals are taught to use this strategy in a systematic way, as Victoria Bell has shown in unpublished studies in the author's laboratory, their reasoning is both faster and more accurate.

Participants mix strategies, and switch from one to another. Sometimes a switch occurs in the middle of a problem; sometimes from one problem to another. There are no fixed sequences of steps that anyone invariably followed. But, regardless of strategy, as a further study showed, inferences that call for only one mental model are easier than those that call for two mental models, which in turn are easier than those that call for three mental models (Van der Henst et al., 2002). Reasoners also develop diverse strategies for reasoning about relations such as "taller than" (Goodwin & Johnson-Laird, 2005, 2006; Roberts, 2000), for reasoning from suppositions (Byrne & Handley, 1997), and for reasoning with quantifiers such as "all" and "some" (Bucciarelli & Johnson-Laird, 1999). All the strategies so far observed reflected a reliance on meaning, and they can be explained in terms of models. But individuals who know logic could make a strategic use of formal rules, and one study has detected signs of the development of formal intuitions (Galotti, Baron, & Sabini, 1986).

MEANING, KNOWLEDGE, AND MODULATION

When human beings reason, they take their knowledge into account. As a result, they often go beyond the explicit information given to them. Suppose, for instance, that the following assertion is true:

> Pat listened to a song, or she listened to some music.

From the further premise that Pat didn't listen to a song, you can infer that she listened to some music. In logic, the *form* of this inference is treated as valid, and psychological theories based on formal rules (e.g., Braine & O'Brien, 1998; Rips, 1994) include such a rule:

> A or B.
> Not A.
> Therefore, B.

They also include a similar rule for cases in which the categorical premise is, *not-B*, and it yields the conclusion, *A*. But, suppose instead that Pat didn't listen to any music. Would you infer that she listened to a song? Obviously not. You know that songs are music, and so if Pat didn't listen to music, she didn't listen to a song. That's part of the meaning of the word "song." Your knowledge of the world can have a similar effect. Given the premises:

> Pat listened to the Beatles' *Yellow Submarine* or she listened to some music.
> Pat didn't listen to any music.

You are unlikely to infer that Pat listened to the Beatles' *Yellow Submarine*, because you know that it is a piece of music.

These two examples are instances of what is known as *modulation* (Johnson-Laird & Byrne, 2002). Meaning, reference, or general knowledge blocks the construction of an otherwise feasible model of an assertion. An inclusive disjunction, *A or B*, is normally interpreted as referring to three possibilities, which have these fully explicit models:

```
    A not-B
not-A     B
    A     B
```

But the disjunction, *Pat listened to a song or she listened to some music*, is modulated so that it refers to just two possibilities for Pat's listening:

```
    song music
not-song music
```

Modulation blocks the third possibility in which Pat listened to a song but not to music.

Most investigations of modulation have concerned conditionals (e.g., Quelhas, Johnson-Laird, & Juhos, 2010). The model theory postulates that the core meaning of *If A then B* also corresponds to a logical interpretation that refers to three possibilities:

```
    A     B
not-A     B
not-A not-B
```

But modulation can block any of these models, apart from the possibility of A and B when A may, or may not, occur, to yield various other interpretations (Johnson-Laird & Byrne, 2002). In addition, it can introduce spatial, temporal, or other relations between the situations referred to in the if-clause and the then-clause. (It can also introduce these relations into disjunctions.) These modulations, in turn, affect the inferences that individuals draw from conditionals. Here is an example from Quelhas et al. (2010):

> If Lisa received the money, then she paid Frederico.
> If she paid Frederico, then he bought a new laptop.
> Lisa received the money.
> Did Lisa receive the money before Frederico bought a new laptop?

Most participants responded, "yes," evidently inferring that the if-clauses in the two premises refer to events that preceded those referred to in the then-clauses. Here is a contrasting example:

> If Tania gave Mauro a scooter, then he did well on the exams.
> If he did well on the exams, then he studied a lot.
> Tania gave Mauro a scooter.
> Did Tania gave Mauro a scooter after he studied a lot?

Again, most participants responded, "yes," but now they evidently inferred that the if-clauses in the two premises referred to events that came after those referred to in the then-clauses. Many studies have shown that models are used to represent explicit spatial and temporal relations (e.g., Byrne & Johnson-Laird, 1989; Carreiras & Santamaria, 1997; Schaeken & Johnson-Laird, 2000; Schaeken, Johnson-Laird, & d'Ydewalle, 1996; Vandierendonck & De Vooght, 1996). But the studies discussed earlier show that individuals use their knowledge to infer temporal relations that are only implicit in the premises.

The potential for meaning, reference, and knowledge to modulate the interpretation of sentential connectives, such as "or" and "if," implies that the system for interpreting sentences must take these factors into account—even if, in the end, a sentence turns out to receive only a logical interpretation. It follows that the interpretative system for sentential connectives cannot work in the "truth functional" way of logic, which takes into account only the truth values of clauses (see, e.g., Jeffrey, 1981). The process of human interpretation is never purely logical: Modulation can add spatial and temporal relations between the events described in a sentence. Hence, sentences of a given grammatical form, such as conditionals or disjunctions, have an indefinite number of different interpretations (pace Evans & Over, 2004).

Inductive Reasoning

Deduction comes with the guarantee that the conclusion of a valid inference must be true if its premises are true. Induction has no such guarantee. Many of the inferences that you make in daily life are inductive—you aim for truth but may miss the target even if your premises are true. For instance, when the starter doesn't turn over your car's engine, your immediate thought is that the battery is dead. You are likely to be right, but there is no guarantee. Likewise, when the car ferry, *Herald of Free Enterprise*, sailed from Zeebrugge on March 6, 1987, its master made the plausible induction that the bow doors had been closed. They had always been closed in the past, and there was no evidence to the contrary. But the doors had *not* been closed, the sea rushed in and the vessel capsized, and over a hundred people drowned. Induction is a risky business. A corollary is that it rules out possibilities over and above those that the premises rule out. It does so because it relies on knowledge, and knowledge is fallible.

Induction is a source of propositions about specific events, such as the closing of the bow doors, and a source of generalizations, such as that car ferries put out to sea with their bow doors closed. And, most important, it is source of explanations. All inductions depend on knowledge and on various constraints, such as its availability (Tversky & Kahneman, 1973), the need for informative hypotheses consistent with the facts, and the similarity of one situation to others (see Johnson-Laird, 2006, Ch. 13). In logic, when a conclusion follows validly from premises, no subsequent information can invalidate it. As new premises are added to existing ones, increasing numbers of logical conclusions therefore follow. Logic is thus "monotonic." But, in daily life, you often withdraw conclusions in the light of subsequent information. Your inferences are "nonmonotonic." Sometimes,you withdraw a conclusion because it was based on an assumption that you made by default, for example, millionaires are right-wing. You encounter a politician who is a millionaire, and so you infer that she is right-wing. But then you learn that she's a Democrat, and so you withdraw your conclusion. The model theory allows for the withdrawal of the consequences of default assumptions. Indeed, this process is an integral part of reasoning based on models (Johnson-Laird & Byrne, 1991). On the one hand, the failure to find a model that serves as a counterexample to a conclusion implies that its inference is valid. On the other hand, the failure to find a model that is consistent with a conclusion—by overturning, say, an assumption made by default—implies that the conclusion is inconsistent with the premises.

Reasoning to Consistency

Many inferences in daily life lead to conflicts with reality. Suppose you know, for example:

> If Ann has gone to get the car, then she will return in 5 minutes.
> Ann has gone to get the car.

You deduce that Ann will return in 5 minutes. In fact, Ann does not return, not even in 20 minutes. You are in a typical situation in which there is a conflict between the consequences of your beliefs and an incontrovertible fact. Something has to "give." At the very least, you have to withdraw your conclusion. You also have to modify your beliefs, but in what way? Should you cease to believe that Ann went to get the car, or that if she did she will return in 5 minutes, or both? Researchers in artificial intelligence have developed various systems of nonmonotonic reasoning to try to deal with such cases (see, e.g., Brewka, Dix, & Konolige, 1997), but psychologists have lagged behind in their investigations of the process. At its heart, there appears to be the creation of diagnostic explanations. You try to imagine a scenario that explains why Ann is not back in 5 minutes. Reasoning that leads in this way from inconsistency to consistency calls for the detection of an inconsistency, the creation of an explanation that accounts for its origins, and perhaps the revision of beliefs (Johnson-Laird, Girotto, & Legrenzi, 2004). But evidence strongly suggests that naïve individuals tend to seek explanations, which as a by-product lead to the revision of their beliefs. The rest of this section accordingly focuses on the discovery of inconsistencies and the creation of explanations that resolve them.

The Discovery of Inconsistencies

A set of propositions is consistent if at least one possibility exists in which they are all true, and it is inconsistent if no such possibility exists. Hence, there is a close relation between consistency and deduction: An inference is valid if the negation of its conclusion is inconsistent with the premises. Inconsistency in a set of propositions implies that at least one proposition in the set is false, and so it is a serious matter in daily life. Sometimes individuals have a plausible model of the world, which turns out to be inconsistent with the facts of the matter, and as a result a disaster occurs, such as a collision at sea (Perrow, 1984, p. 230). The ability to detect inconsistencies is accordingly central to rationality.

You could use logic to detect an inconsistency in a set of propositions, but the method is psychologically implausible: You are supposed to select a proposition from the set and try to prove its negation from the remaining propositions. If you succeed, then the original set is inconsistent; otherwise, it is consistent. It follows that inconsistency should be easier to establish than consistency: With an inconsistency, you can stop as soon as you have proved the negated proposition, but with consistency you must go on searching until you have exhausted all possible proofs (or yourself). But, however you seek to assess consistency, the task is computationally intractable. The demands it places on time and memory increase at such a rate as the size of the set of propositions increases that the task soon defeats any feasible computational system. The question remains, however: Even with a small set of propositions, how do you assess their consistency?

The model theory provides this answer: Individuals evaluate the consistency of a set of propositions by searching for a model of a possibility in which all the propositions are true. If they find such a model, the propositions are consistent; otherwise, they are inconsistent (Johnson-Laird, Legrenzi, Girotto, & Legrenzi, 2000). Hence, contrary to the logical account given earlier, consistency should be easier to establish than inconsistency, because you can stop as soon as you have found one possibility in which all the propositions hold, whereas with an inconsistency you have to examine the possibilities exhaustively in order to establish that not one exists in which all the propositions hold.

Experiments have supported the model theory. Consider, for instance, whether these propositions about what is on a table could all be true at the same time:

> If there isn't an apple then there is a banana.
> If there is a banana then there is a cherry.
> There isn't an apple and there is a cherry.

You are likely to begin by considering an obvious possibility for the first proposition, which corresponds to its one explicit mental model (see earlier):

> not-apple banana.

This possibility fits the second proposition, which adds a further fruit on top of the table:

> not-apple banana cherry

The third proposition holds in this model, and so you infer that the set of propositions is consistent. In contrast, consider this description:

> There is an apple or there is a banana.
> There isn't a banana or there is a cherry.
> There isn't an apple and there is a cherry.

You begin by considering the obvious possibility for the first proposition, corresponding to its first mental model:

> apple

You update this model according to the second proposition:

> apple not-banana

But this possibility is not consistent with the third proposition, and so you have to retrace your steps. At length, you discover a possibility in which all three propositions hold:

> not-apple banana cherry

But this sort of problem in which your initial model leads you astray should be harder than the first sort of problem. An experiment in which the participants were over 500 of the best high school graduates in

Italy showed that the first sort of problem had a robust advantage in accuracy (of 15%) over the second sort of problem—even when conditionals and disjunctions were counterbalanced. Likewise, as the model theory predicts, the consistent problems were easier than the inconsistent problems.

The principle of truth is central to the model theory, because it predicts the occurrence of illusory inferences. Consider this description:

> The tray is portable or else not both beautiful and heavy.
> The tray is portable and not beautiful.

The first mental model of the tray according to the disjunction is:

> portable

The tray lacks the property of being beautiful, and so individuals should judge that the second assertion is consistent with it. They would be wrong. If the tray is portable, the first clause of the disjunction is true and so its second clause is false, that is, the tray *is* both beautiful and heavy. This tray is not consistent with the second assertion in the description. But if it is false that the tray is portable, then the tray is also inconsistent with the second assertion. Hence, the two assertions are inconsistent. An experiment compared the illusory problems with similar control problems, and it corroborated the theory's predictions. The participants responded more accurately to the control problems (86% correct) than to the illusory problems (27% correct), and only 11 of 459 participants went against this trend (Legrenzi, Girotto, & Johnson-Laird, 2003). A further experiment conveyed the meaning of "or else" with an unambiguous statement of an exclusive disjunction: "Only one of the following assertions is true." Once again, the participants succumbed to illusions, but they responded correctly to control problems.

The Creation of Explanations

Reasoning in daily life often calls for the generation of explanations, especially when you have detected an inconsistency among the propositions that you believe. For example, in the case of Ann's failure to return in 5 minutes when she went to get the car, you try to make explanatory inductions about what may have happened:

> Perhaps the battery was dead and she couldn't get the car to start.
> Perhaps she didn't remember where we were and got lost on the way.

You use your knowledge and any relevant evidence to generate possibilities. Human reasoners easily outperform any current computer program in envisaging putative explanations. Given two sentences selected at random from different stories, such as:

> Celia made her way to a shop that sold TV sets.
> She had recently had her ears pierced.

they readily offer explanations of what's going on, such as: Celia was getting reception in her ears and wanted the TV shop to investigate, or Celia had bought some new earrings and wanted to see how they looked on closed-circuit TV (Johnson-Laird, 2006, Ch. 14). This ability to create explanations underlies both science and superstition. The difference is that scientists test their explanations.

When you discover an inconsistency, you try to frame a causal explanation that accounts for its origin. The model theory postulates that the basic unit of explanations is a cause and its effect, in which the effect resolves the inconsistency. It makes possible the facts of the matter, and it repudiates at least one of your previous premises, which you then take to refer to a counterfactual possibility, that is, a situation that was once possible but that did not occur (see Byrne, 2002, 2005; Quelhas & Byrne, 2003). According to the model theory, the *meaning* of a causal relation between two states of affairs, a cause and its effect, refers to what is possible and what is impossible in their co-occurrences. The claim is controversial, but it has been corroborated experimentally (Goldvarg & Johnson-Laird, 2001). In daily life, the normal constraint is that an effect does not precede its cause in time (see, e.g., Tversky & Kahneman, 1982). Hence, the theory adopts this constraint. A computer program implements this account for simple inconsistencies, such as:

> If the trigger is pulled then the pistol will fire.
> The trigger is pulled.
> But the pistol does not fire. Why not?

The program constructs mental models of the premises, detects the inconsistency with the facts, and uses its knowledge base of explicit models of causal relations to construct a causal chain that resolves the inconsistency, for example, a person emptied the pistol and so there were no bullets in the pistol (Johnson-Laird et al., 2004). When individuals were given 20 different inconsistencies, such as this example about the pistol, but from varied domains, they were usually able to create a causal explanation (Johnson-Laird et al., 2004). Most of these explanations repudiated the conditional. In two further experiments with the scenarios, the participants rated the statements of a cause and its effect as the most probable explanations, for example: A prudent person had unloaded the pistol and there were no bullets in the chamber. The cause alone was rated as less probable, but as more probable than the effect alone, which in turn was rated as more probable than an explanation that repudiated the categorical premise, for example, the trigger wasn't really pulled. The greater probability assigned to the conjunction of the cause and effect than to either of its clauses is an instance of the "conjunction" fallacy in which a conjunction is wrongly judged to be more probable than its constituents (Tversky & Kahneman, 1983). Recent studies have similarly shown that participants rate such explanations as more probable than simple denials of either the conditional premise or the categorical premise (Khemlani & Johnson-Laird, 2011).

In sum, reasoners can resolve inconsistencies. They use their knowledge to try to create a causal model that makes sense of the facts. Their reasoning may resolve the inconsistency or fail to yield any explanation whatsoever. One view of rational changes to beliefs is that they should incorporate the facts with minimal changes. As William James (1907, p. 59) wrote: "[The new fact] preserves the older stock of truths with a minimum of modification, stretching them just enough to make them admit the novelty." Such parsimony is sensible, and many cognitive scientists have advocated minimalism both for science and for daily life (e.g., Gärdenfors, 1992; Harman, 1986). Likewise, computer programs for artificial intelligence have modeled minimal changes (e.g., deKleer, 1986), and measures have been developed to calculate what counts as a minimal change (Elio & Pelletier, 1997; Harman, 1986). What the results reviewed in this section show is that naïve individuals are happy to sacrifice minimalism in the cause of an explanation (see also Walsh & Johnson-Laird, 2009).

Conclusions

What is a mental model? The answer to this question conveys the main points of this chapter. A mental model is a representation of the world that is constructed from perception, memory, or imagination, and that underlies thinking. Three key properties distinguish a mental model from other proposed sorts of mental representation:

1. A mental model represents a possibility: It is a conjunction of entities, their properties, and interrelations. Strictly speaking, a mental model of a situation captures what is common to the different ways in which it could occur. Hence, a description that makes explicit several alternative possibilities has models corresponding to each of them. As a result, you have greater difficulty in envisaging the description and in reasoning from it than from a description that yields only a single mental model. You infer that a conclusion that holds in all models is necessary given the description, one that holds in most models is probable, and one that holds in at least one model is possible. And you can refute a putative conclusion by discovering a model that is a counterexample to it.
2. A mental model is iconic insofar as it can be, which is to say that its structure corresponds to the structure of what it represents unlike, say, the syntactic structure of a sentence. Visual images are iconic, but mental models can also contain symbolic elements, such as negation (see Schroyens, Schaeken, & d'Ydewalle, 2001). And they can represent properties and relations that have meanings that cannot be visualized. As a corollary, models differ from propositional representations, which are syntactically structured representations in a mental language. The symbolic components of models also distinguish them from putative representations rooted in a sensory modality.
3. A mental model represents what is true as opposed to what is false. This principle of truth enables models to be much more parsimonious than, say, truth tables, which represent both what is true and what is false. As a result, models put much less of a load on the processing capacity of working memory. But sometimes you err as a result. In the case of those inferences for which falsity matters, you are likely to succumb to the illusion that a conclusion is valid when in fact it is not, and vice versa. Suppose, for instance, you know that either Pat called her mother on Monday or otherwise she went to see her mother on Tuesday or else on Wednesday but not both days. You are likely to think of these as three alternative possibilities. So, given, say, that she went to see her mother on Tuesday, you are likely to infer that she didn't go to see her on Wednesday. But suppose that Pat called her mother on Monday. The first clause of the principal exclusive disjunction is true, and so it is false that she went to see her mother either on Tuesday or Wednesday, but not both. And one way in which it could be false is that she went to see her mother on both days. Your inference isn't valid, even though it is compelling.

Finally, mental models have been proposed for domains remote from mainstream cognitive psychology. Bowlby (1988), for example, argued that models of caregivers play a crucial part in the development of children. Models also appear to underlie the reasoning of individuals suffering from psychological illnesses, and their reasoning—contrary to an assumption of cognitive therapy (e.g., Beck, 1976)—is superior to the reasoning of nonclinical controls, though only on topics relating to the patients' illnesses (Johnson-Laird, Mancini, & Gangemi, 2006). Models also predict an effect of personality on reasoning: Individuals who are open to experience tend to think of possibilities outside the premises and therefore they tend to make inductions, whereas those with the mirror-image traits tend to stick to the possibilities to which the premises refer and therefore they tend to make deductions (Fumero, Santamaría, & Johnson-Laird, 2010).

Future Directions

1. How might mental models underlie the mental representation of stereotypes and prototypes?
2. Models appear to underlie "extensional" estimates of probability based on the different possible ways in which an event might occur, but what role, if any, do they play in "intensional" estimates based on intuitions about evidence?
3. How do children develop the ability to construct and to manipulate mental models?

Acknowledgments

The theory of mental models has developed as a result of the work of many researchers, too numerous to list here, but readers should consult the references for the names of many of them. Much of the research reported in the chapter was made possible by grants from the National Science Foundation (BCS 0076287 and SES 0844851) to study strategies in reasoning, and deductive and probabilistic reasoning.

References

Almor, A. (1999). Noun-phrase anaphora and focus: The informational load hypothesis. *Psychological Review*, *106*, 748–765.

Barrouillet, P., & Lecas, J-F. (2000). Illusory inferences from a disjunction of conditionals: A new mental models account. *Cognition*, *76*, 3–9.

Barsalou, L.W. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences*, *22*, 577–660.

Barwise, J. (1987). *The situation in logic*. Stanford, CA: CSLI.

Bauer, M. I., & Johnson-Laird, P. N. (1993). How diagrams can improve reasoning. *Psychological Science*, *4*, 372–378

Beck, A. T. (1976). *Cognitive therapy and the emotional disorders*. New York: Meridian.

Bell, V., & Johnson-Laird, P. N. (1998). A model theory of modal reasoning. *Cognitive Science*, *22*, 25–51.

Biederman, I. (1995). Visual object recognition. In S. M. Kosslyn & D. N. Osherson (Eds.), *An invitation to cognitive science, Vol. 2. Visual cognition* (2nd ed., pp. 121–165). Cambridge, MA: MIT Press.

Bowlby, J. (1988). *A secure base: Clinical applications of attachment theory*. London: Routledge.

Braine, M. D. S., & O'Brien, D. P. (Eds.). (1998). *Mental logic*. Mahwah, NJ: Erlbaum.

Bransford, J. D., Barclay, J. R., & Franks, J. J. (1972). Sentence memory: A constructive versus an interpretive approach. *Cognitive Psychology*, *3*, 193–209.

Brewka, G., Dix, J., & Konolige, K. (1997). *Nonmonotonic reasoning: An overview*. Stanford, CA: CLSI, Stanford University.

Bruner, J. S., Goodnow, J. S., & Austin, G. G. (1956). *A study of thinking*. New York: Wiley.

Bucciarelli, M., & Johnson-Laird, P. N. (1999). Strategies in syllogistic reasoning. *Cognitive Science*, *23*, 247–303.

Bucciarelli, M., & Johnson-Laird, P. N. (2005). Naïve deontics: A theory of meaning, representation, and reasoning. *Cognitive Psychology*, *50*, 159–193.

Byrne, R. M. J. (2002). Mental models and counterfactual thoughts about what might have been. *Trends in Cognitive Sciences*, *6*, 426–431

Byrne, R. M. J. (2005). *The rational imagination: How people create alternatives to reality*. Cambridge, MA: MIT Press.

Byrne, R. M. J., & Handley, S. J. (1997). Reasoning strategies for suppositional deductions. *Cognition*, *62*, 1–49

Byrne, R. M. J., & Johnson-Laird, P. N. (1989). Spatial reasoning. *Journal of Memory and Language*, *28*, 564–575.

Carreiras, M., &Santamaria, C. (1997). Reasoning about relations: Spatial and nonspatial problems. *Thinking and Reasoning*, *3*, 191–208.

Cowles, W., & Garnham, A. (2005). Antecedent focus and conceptual distance effects in category noun-phrase anaphora. *Language and Cognitive Processes*, *20*, 725–750.

Craik, K. (1943). *The nature of explanation*. Cambridge, England: Cambridge University Press.

de Kleer, J. (1986). An assumption-based TMS. *Artificial Intelligence*, *28*, 127–162

Elio, R., & Pelletier, F. J. (1997). Belief change as prepositional update. *Cognitive Science*, *21*, 419–460.

Evans, J.St. B. T. (2003). In two minds: Dual process accounts of reasoning. *Trends in Cognitive Sciences*, *7*, 454–459.

Evans, J. St. B. T., & Over, D. E. (1996). *Rationality and reasoning*. Hove, England: Psychology Press.

Evans, J. St. B. T., & Over, D. E. (2004). *If*. Oxford, England: Oxford University Press.

Feldman, J. (2000). Minimization of Boolean complexity in human concept learning. *Nature*, *407*, 630–633.

Fumero, A., Santamaría, C., & Johnson-Laird, P. N. (2010). Reasoning and autobiographical memory for personality. *Experimental Psychology*, *57*, 215–220.

Galotti, K. M., Baron, J., & Sabini, J. P. (1986). Individual differences in syllogistic reasoning: Deduction rules or mental models? *Journal of Experimental Psychology: General*, *115*, 16–25.

García-Madruga, J. A., Moreno, S., Carriedo, N., Gutiérrez, F., & Johnson-Laird, P. N. (2001). Are conjunctive inferences easier than disjunctive inferences? A comparison of rules and models. *Quarterly Journal of Experimental Psychology*, *54A*, 613–632.

Gärdenfors, P. (1992). Belief revision: An introduction. In P. Gärdenfors (Ed.), *Belief revision* (pp. 1–20). Cambridge, England: Cambridge University Press.

Garnham, A. (1987). *Mental models as representations of discourse and text*. Chichester, England: Ellis Horwood.

Garnham, A. (2001). *Mental models and the interpretation of anaphora*. Hove, England: Psychology Press.

Gentner, D., & Stevens, A. L. (Eds.). (1983). *Mental models*. Hillsdale, NJ: Erlbaum.

Gernsbacher, M. A. (1990). *Language comprehension as structure building*. Hillsdale, NJ: Erlbaum.

Girotto, V., Mazzocco, A., & Tasso. A. (1997). The effect of premise order in conditional reasoning: a test of the mental model theory. *Cognition*, *63*, 1–28.

Glenberg, A. M., Meyer, M., & Lindem, K. (1987). Mental models contribute to foregrounding during text comprehension. *Memory and Language*, *26*, 69–83.

Goldvarg, Y., & Johnson-Laird, P. N. (2000). Illusions in modal reasoning. *Memory and Cognition*, *28*, 282–294.

Goldvarg, Y., & Johnson-Laird, P. N. (2001). Naïve causality: A mental model theory of causal meaning and reasoning. *Cognitive Science*, *25*, 565–610.

Goodwin, G., & Johnson-Laird, P. N. (2005). Reasoning about relations. *Psychological Review*, *112*, 468–493.

Goodwin, G., & Johnson-Laird, P. N. (2006). Reasoning about the relations between relations. *Quarterly Journal of Experimental Psychology*, *59*, 1047–1069.

Goodwin, G., & Johnson-Laird, P. N. (2010). Conceptual illusions. *Cognition*, *114*, 253–265.

Goodwin, G. P., & Johnson-Laird, P. N. (2011). Mental models of Boolean concepts. *Cognitive Psychology*, *63*, 34–59.

Harman, G. (1986). *Change in view: Principles of reasoning*. Cambridge, MA: MIT Press.

Hegarty, M. (1992). Mental animation: Inferring motion from static diagrams of mechanical systems. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *18*, 1084–1102.

Hunt, E. B. (1962). *Concept learning: An information processing problem*. New York: Wiley.

Inhelder, B., & Piaget, J. (1958). *The growth of logical thinking from childhood to adolescence*. London: Routledge & Kegan Paul.

James, W. (1907). *Pragmatism—A new name for some old ways of thinking*. New York: Longmans, Green.

Jeffrey, R. (1981). *Formal logic: Its scope and limits* (2nd ed.). New York: McGraw-Hill.

Johnson-Laird, P. N. (1983). *Mental models*. Cambridge, MA: Harvard University Press.

Johnson-Laird, P. N. (2006). *How we reason*. Oxford, England: Oxford University Press.

Johnson-Laird, P. N., & Byrne, R. M. J. (1991). *Deduction*. Hillsdale, NJ: Erlbaum.

Johnson-Laird, P. N., & Byrne, R. M. J. (2002). Conditionals: A theory of meaning, pragmatics, and inference. *Psychological Review*, *109*, 646–678.

Johnson-Laird, P. N., Byrne, R. M. J., & Schaeken, W. S. (1992). Propositional reasoning by model. *Psychological Review*, *99*, 418–439.

Johnson-Laird, P. N., Girotto, V., & Legrenzi, P. (2004). Reasoning from inconsistency to consistency. *Psychological Review*, *111*, 640–661.

Johnson-Laird, P. N., & Hasson, U. (2003). Counterexamples in sentential reasoning. *Memory and Cognition*, *31*, 1105–1113.

Johnson-Laird, P. N., Legrenzi, P., Girotto, P., & Legrenzi, M. S. (2000). Illusions in reasoning about consistency. *Science*, *288*, 531–532.

Johnson-Laird, P. N., Legrenzi, P., Girotto, V., Legrenzi, M., & Caverni, J-P. (1999). Naïve probability: A mental model theory of extensional reasoning. *Psychological Review*, *106*, 62–88.

Johnson-Laird, P. N., Mancini, F., & Gangemi, A. (2006). A hyper emotion theory of psychological illnesses. *Psychological Review*, *113*, 822–841.

Johnson-Laird, P. N., & Savary, F. (1999). Illusory inferences: A novel class of erroneous deductions. *Cognition*, *71*, 191–229.

Johnson-Laird, P. N., & Wason, P. C. (Eds.). (1976). *Thinking: Readings in cognitive science*. Cambridge, England: Cambridge University Press.

Kahneman, D., & Frederick, S. (2002). Representativeness revisited: Attribute substitution in intuitive judgement. In T. Gilovich, D. Griffin, & D. Kahneman (Eds.), *Heuristics and biases: The psychology of intuitive judgement* (pp. 49–81.) Cambridge, England: Cambridge University Press.

Kamp, H., & Reyle, U. (1993). *From discourse to logic*. Dordrecht, The Netherlands: Kluwer.

Kemp, C., & Tenenbaum, J. B. (2008). The discovery of structural form. *Proceedings of the National Academy of Sciences USA*, *105*, 10687–10692.

Khemlani, S., & Johnson-Laird, P. N. (2009). Disjunctive illusory inferences and how to eliminate them. *Memory and Cognition*, *37*, 615–623.

Khemlani, S., & Johnson-Laird, P. N. (2011). The need to explain. *Quarterly Journal of Experimental Psychology*, *64*, 276–288.

Kintsch, W. (1988). The role of knowledge in discourse comprehension: A construction-integration model. *Psychological Review*, *95*, 163–182.

Knauff, M., Fangmeier, T., Ruff, C. C., &Johnson-Laird, P. N. (2003). Reasoning, models, and images: Behavioral measures and cortical activity. *Journal of Cognitive Neuroscience*, *4*, 559–573.

Knauff, M., & Johnson-Laird, P. N. (2002). Imagery can impede inference. *Memory and Cognition*, *30*, 363–371.

Kripke, S. (1963). Semantical considerations on modal logic. *Acta Philosophica Fennica*, *16*, 83–94.

Kroger, J. K., Nystrom, L. E., Cohen, J. D., & Johnson-Laird, P. N. (2008). Distinct neural substrates for deductive and mathematical processing. *Brain Research*, *1243*, 86–103.

Lee, N. Y. L., Goodwin, G. P., & Johnson-Laird, P. N. (2008). The psychological problem of Sudoku. *Thinking and Reasoning*, *14*, 342–364.

Legrenzi, P., Girotto, V., & Johnson-Laird, P. N. (2003). Models of consistency. *Psychological Science*, *14*, 131–137.

Magliano, J. P., Miller, J., & Zwaan, R. A. (2001). Indexing space and time in film understanding. *Applied Cognitive Psychology*, *15*, 533–545.

Mar, R. A., & Oatley, K. (2008). The function of fiction is the abstraction and simulation of social experience. *Perspectives on Psychological Science*, *3*, 173–192.

Markman, A. B., & Dietrich, E. (2000). Extending the classical view of representation. *Trends in Cognitive Science*, *4*, 470–475.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. San Francisco: W. H. Freeman.

Medin, D. L., Lynch, E. B., & Solomon, K. O. (2000). Are there kinds of concepts? *Annual Review of Psychology*, *51*, 121–147.

Medin, D. L., & Smith, E. E. (1984). Concepts and concept formation. *Annual Review of Psychology*, *35*, 113–138.

Metzler, J., & Shepard, R. N. (1982). Transformational studies of the internal representations of three-dimensional objects. In R. N. Shepard & L. A. Cooper *Mental images and their transformations* (pp. 25–71.) Cambridge, MA: MIT Press.

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

Nosofsky, R. M., Palmeri, T. J., & Mc Kinley, S. C. (1994). Rule-plus-exception model of classification learning. *Psychological Review*, *101*, 53–79.

Oberauer, K. (2006). Reasoning with conditionals: A test of formal models of four theories. *Cognitive Psychology*, *53*, 238–283.

Ormerod, T. C., & Richardson, J. (2003). On the generation and evaluation of inferences from single premises. *Memory and Cognition*, *31*, 467–478.

Peirce, C. S. (1931–1958). *Collected papers of Charles Sanders Peirce* (C. Hartshorne, P. Weiss, & A. Burks, Eds.). Cambridge, MA: Harvard University Press.

Perrow, C. (1984). *Normal accidents: Living with high-risk technologies*. New York: Basic Books.

Pylyshyn, Z. (2003). Return of the mental image: Are there really pictures in the head? *Trends in Cognitive Science*, *7*, 113–118.

Quelhas, A. C., & Byrne, R. M. J. (2003). Reasoning with deontic and counterfactual conditionals. *Thinking and Reasoning*, *9*, 43–66.

Quelhas, A. C., Johnson-Laird, P. N., & Juhos, C. (2010). The modulation of conditional assertions and its effects on reasoning. *Quarterly Journal of Experimental Psychology*, *63*, 1716–1739.

Rader, A.W., & Sloutsky, V. M. (2002). Processing of logically valid and logically invalid conditional inferences in discourse comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*, 59–68.

Radvansky, G. A., & Copeland, D. E. (2006). Walking through doorways causes forgetting: Situation models and experienced space. *Memory and Cognition*, *34*, 1150–1156.

Rinck, M., & Bower, G. (1995). Anaphor resolution and the focus of attention in situation models. *Memory and Language*, *34*, 110–131.

Rips, L. J. (1994). *The psychology of proof*. Cambridge, MA: MIT Press.

Roberts, M. J. (2000). Strategies in relational inference. *Thinking and Reasoning*, *6*, 1–26.

Santamaría, C., & Johnson-Laird, P. N. (2000). An antidote to illusory inferences. *Thinking and Reasoning*, *6*, 313–333.

Schaeken, W., & Johnson-Laird, P. N. (2000). Strategies in temporal reasoning. *Thinking and Reasoning*, *6*, 193–219.

Schaeken, W. S., Johnson-Laird, P. N., & d' Ydewalle, G. (1996). Mental models and temporal reasoning. *Cognition*, *60*, 205–234.

Schroyens, W., Schaeken, W., & Handley, S. (2003). In search of counter examples: Deductive rationality in human reasoning. *Quarterly Journal of Experimental Psychology*, *56A*, 1129–1145.

Schroyens, W., Schaeken, W., & d'Ydewalle, G. (2001). The processing of negations in conditional reasoning: A meta-analytic case study in mental model and/or mental logic theory. *Thinking and Reasoning*, *7*, 121–172.

Shepard, R. N., Hovland, C. I., & Jenkins, H. M. (1961). Learning and memorization of classifications. *Psychological Monographs: General and Applied*, *75*, 1–42.

Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin*, *119*, 3–22.

Stanovich, K. E. (1999). *Who is rational? Studies of individual differences in reasoning*. Mahwah, NJ: Erlbaum.

Stevenson, R. J., Nelson, A. W. R., & Stenning, K. (1995). The role of parallelism in strategies of pronoun comprehension. *Language and Speech*, *38*, 393–418.

Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology*, *5*, 207–232.

Tversky, A., & Kahneman, D. (1982). Causal schemas in judgements under uncertainty. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgement under uncertainty: Heuristics and biases*. (pp. 117–128). Cambridge, England: Cambridge University Press.

Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, *90*, 292–315.

Ullman, S. (1996). *High-level vision: Object recognition and visual cognition*. Cambridge, MA: MIT Press.

Van der Henst, J-B., Yang, Y., & Johnson-Laird, P. N. (2002). Strategies in sentential reasoning. *Cognitive Science*, *26*, 425–468.

Vandierendonck, A., & De Vooght, G. (1996). Evidence for mental-model based reasoning: A comparison of reasoning with time and space concepts. *Thinking and Reasoning*, *2*, 249–272.

Van Dijk, T. A., & Kintsch, W. (1983). *Strategies of discourse comprehension*. New York: Academic Press.

Verschueren, N., Schaeken, W., & d' Ydewalle, G. (2005). A dual-process specification of causal conditional reasoning. *Thinking and Reasoning*, *11*, 278–293.

Vigo, R. (2009). Modal similarity. *Journal of Experimental and Theoretical Artificial Intelligence*, iFirst, 1–16.

Walsh, C., & Johnson-Laird, P. N. (2004). Co-reference and reasoning. *Memory and Cognition*, *32*, 96–106.

Walsh, C. R., & Johnson-Laird, P. N. (2009). Changing your mind. *Memory and Cognition*, *37*, 624–631.

Wason, P. C., & Evans, J. St. B. T. (1975). Dual processes in reasoning? *Cognition*, *3*, 141–154.

Yang, Y., & Johnson-Laird, P. N. (2000). How to eliminate illusions in quantified reasoning. *Memory and Cognition*, *28*, 1050–1059.

Further Reading

Byrne, R. M. J. (2005). *The rational imagination: How people create alternatives to reality*. Cambridge, MA: MIT Press.

Johnson-Laird, P. N. (2006). *How we reason*. Oxford, England: Oxford University Press.

Khemlani, S., & Johnson-Laird, P. N. (2012). The processes of inference. *Argument and Computation*, 1–17, iFirst.