

12

Mental Models

P. N. Johnson-Laird

The modern formulation of the concept of a mental model is due to Kenneth Craik (1943). In his remarkably prescient book, *The Nature of Explanation*, he argued that human beings translate external events into internal models and reason by manipulating these symbolic representations. They can translate the resulting symbols back into actions or recognize a correspondence between them and external events. Although Craik died before the invention of the programmable digital computer, he would have recognized the computer metaphor. He argued that the physical substrate of the brain is less pertinent than the way it functions. He wrote (Craik 1943, p. 51), "By a model we thus mean any physical or chemical system which has a similar relation-structure to that of the processes it imitates. By 'relation-structure' I do not mean some obscure physical entity which attends the model, but the fact that it is a physical working model which works in the same way as the processes it parallels, in the aspects under consideration at any moment. . . ." Hence for Craik a mental model was preeminently a dynamic representation or simulation of the world. He had little to say about the form of such representations or about the processes that manipulate them. It was bold enough to postulate representations in the heyday of Behaviorism.

Most cognitive scientists following Craik have adopted the basic tenet that the mind is a symbolic system (see, for example, Miller, Galanter, and Pribram 1960, Newell and Simon 1976). And during the 1980s there has been an enormous growth in studies of mental models. These studies are so extraordinarily diverse, however, that they seem to have little in common beyond the bare appeal to symbolic representations of some sort. Thus explanations of visual perception, the comprehension of discourse, reasoning, and the representation of knowledge and expertise have all invoked versions of the mental-model hypothesis. My aim in this chapter is to bring some order to this diversity. Although many writers have gone out of their way to draw distinctions between alternative concepts of mental models, the theories may differ more than they ought to because they nearly all concern the same underlying reality. It is as though explorers keep reporting the existence of a

hitherto unknown animal, but their fragmentary glimpses of it convince them that they are observing different creatures.

I begin by sketching the role of models in perception and discourse, but only briefly to avoid trespassing on other chapters in this volume. Next I consider reasoning as a process of manipulating models and then their role as representations of knowledge. Finally I present a critique of the theory of mental models and outline some of the major problems in its development.

12.1 Perception as a Source of Mental Models

A primary source of mental representations is perception. As I argue in my book on mental models (Johnson-Laird 1983), the simplest creatures, such as single-cell organisms, merely react physically to their immediate environment. Paramecia, for instance, bump into an obstacle, and the ensuing chemical changes cause their cilia to beat in the opposite direction. But evolution has produced creatures with a nervous system that detects energy from distant objects and that directly modulates the activity of the neurons controlling stereotyped responses. More complicated creatures use the information impinging on the sensorium to compute an internal representation that in turn is used by the processes controlling action. These internal representations may encode relatively superficial features of the world. The housefly's visual system, for example, controls its flight pattern by way of a representation that probably makes explicit only certain features of the visual field, such as a rapid expansion as the fly approaches a surface (see Reichardt and Poggio 1981). Human vision, as David Marr and his colleagues have emphasized, depends on the construction of a series of symbolic representations culminating in a three-dimensional *model* of the spatial relations among objects (see Marr 1982 and chapter 15). This model makes explicit *what is where* to our conscious processes of judgment and thereby enables us to navigate our way through the world avoiding obstacles and hazards.

Our models need to integrate the information from all the senses and from general knowledge—the sights, sounds, smells, and possibilities of the world. Our capacity to envisage different situations appears to be limitless, but the brain cannot contain an infinite number of preexisting symbols no more than a library can contain an infinite number of books. The vast range of mental models must be constructed out of finite means—out of primitive symbols and the basic processes that operate on them. In the case of vision, computational procedures carried out by the brain convert the retinas' response to light into a model of the objects reflecting that light, and these procedures embody constraints based on the nature of the world (Marr 1982). Our phenomenological experience of the world is a triumph of natural selection. We seem to perceive the world directly, not a representation of it. Yet this

phenomenology is illusory: what we perceive depends on both what is in the world and what is in our heads—on what evolution has “wired” into our nervous systems and what we know as a result of experience. The limits of our models are the limits of our world.

12.2 Mental Models of Discourse

Wittgenstein (1922, section 4.01) in his celebrated “picture” theory of meaning wrote, “The proposition is a model of reality as we imagine it.” When I tell you, say, that there is a table in front of the stove in my kitchen, you can imagine the arrangement even if you cannot see it. A major function of language is thus to enable us to experience the world by proxy, because we can envisage how it is on the basis of a verbal description. The assertion “A table is in front of the stove” establishes a relation between two entities. Hence its model contains two mental tokens, corresponding to the table and the stove, interrelated in a way that corresponds to the spatial relation between them. In short, discourse models make explicit the structure *not* of sentences but of situations as we perceive or imagine them (Johnson-Laird 1983, p. 419).

One reason for believing that people construct models is that the hypothesis explains a central feature of comprehension. The explicit content of a discourse is usually only a blueprint for a state of affairs: it relies on the reader or listener to flesh out the missing details. Such “bridging” inferences are rapid and automatic, and people are seldom aware of them though they may show up in recalling the discourse (see Clark 1977). These inferences sometimes depend on general knowledge, as a number of cognitive scientists have argued (for example, Schank and Abelson 1977, Sanford and Garrod 1981). But the key fact is that they yield conclusions of a sort that would be explicit only in models of situations. Bransford and his colleagues have demonstrated this phenomenon in a number of studies. For example, Bransford, Barclay, and Franks (1972) observed that when subjects were presented with the sentence

Three turtles rested on a floating log and a fish swam beneath them.
they later confused it with

Three turtles rested on a floating log and a fish swam beneath it.

If the subjects had imagined the situation described in the original sentence, constructing a model in which the turtles are on the log and the fish swims beneath it, then the model would also represent the fish as swimming beneath the log. This explanation is corroborated by the lack of any such error when the original sentence yields no such model:

Three turtles rested beside a floating log and a fish swam beneath them.

There is also abundant evidence that the coherence of discourse depends in part on how easy it is to construct a single mental model from it (see Garnham, Oakhill, and Johnson-Laird 1982, Ehrlich and Johnson-Laird 1982, Oakhill and Garnham 1985). Kannan Mani and I have similarly shown that passages calling for a single model of a spatial layout are easier to remember than indeterminate descriptions that are consistent with more than one layout (Mani and Johnson-Laird 1982). Thus the description

The spoon is to the left of the knife.

The plate is to the right of the knife.

The fork is in front of the spoon.

The cup is in front of the knife.

describes the following arrangement of objects:

spoon knife plate

fork cup

But a change of one word in the second sentence,

The plate is to the right of the spoon.

yields an indeterminate description consistent with two quite different layouts:

spoon knife plate spoon plate knife

fork cup fork cup

After our subjects had classified a series of determinate and indeterminate descriptions as true or false of diagrams of layouts, they were given an unexpected recognition test of their memory for the descriptions. They retained the gist of the determinate descriptions very much better than the gist of the indeterminate descriptions. Yet they were better at recognizing verbatim details of the indeterminate descriptions. This pattern of results bore out our hypothesis that the subjects would construct a model of a determinate description to compare with a diagram, but try to commit an indeterminate description to memory rather than construct its alternative possible models. It is difficult to see how subjects could even be sensitive to the difference between the two sorts of descriptions unless they attempt to construct models of them.

The existence of two forms of representation—linguistic representations and discourse models—has been corroborated in a number of other experiments. Thus where two expressions with different meanings, such as

the man with the martini

the man standing by the window

occur in contexts in which they refer to the same individual, then as Garnham (1987) has shown, subjects readily confuse the two in an

unexpected recognition test of the passage. But subjects who are warned about the recognition test retain their memory for the particular expressions used to refer to individuals. As Garnham points out, the outcome of this experiment is at odds with the idea that discourse is encoded solely in a semantic network or in any other format that represents merely the meanings of expressions. Referents and, in particular, tokens corresponding to individuals, must be independently represented in a discourse model based on the meanings of the expressions, general knowledge, and the prior context established in the model.

The distinction between linguistic representations and discourse models accounts for a variety of linguistic phenomena, notably anaphoric reference (see Sag and Hankamer 1984 and chapter 11). Linguistic representations are expressions in a mental language (Kintsch 1974, Kintsch and van Dijk 1978, Fodor 1975), which can cope with those anaphora that depend on meanings or forms of words. For example, in the discourse

The cats were biting the dogs.

The fleas were too.

the interpretation of the second elliptical sentence requires access to the surface form of the previous sentence. (Compare its interpretation with the one that would occur if the previous sentence had a different surface form: 'The dogs were being bitten by the cats.' Unless the listener has retained a linguistic representation, these sorts of anaphora are indeed difficult to understand (see Garnham and Oakhill 1987). Other anaphoric expressions, however, refer to entities that have been introduced earlier in the discourse (or by extralinguistic means, such as a gesture). In the discourse

There's a large wooden table in front of the stove.

It has four chairs, one on each side.

the pronoun "it" harks back, not to some form of words but to a particular referent that has been introduced earlier. There is therefore a need to go beyond the linguistic representation of discourse to a model of the situation.

A further reason for postulating discourse models concerns the vexed issue of truth. Theories based solely on linguistic representations do not say anything about how words relate to the world (Johnson-Laird, Herrmann, and Chaffin 1984). Until such relations are established, the question of whether a description is true or false cannot arise. Mental models are symbolic structures, and the relation of a model to the world cannot simply be read off from the model. So how is the truth or falsity of an assertion judged in relation to the world? The answer is that a discourse will be judged true if its mental model can be embedded in

the model of the world. Thus, for example, you will judge my remark about the table being in front of the stove as true if it corresponds to your perception of the world, that is, a model based on the assertion can be embedded within a perceptual model of the situation (Johnson-Laird 1983, pp. 247, 441). (You may be in a position to relate the linguistic representation directly to a perceptual model of the world.) The notion of embedding means that the same individuals with the same properties and relations are preserved from one model to the other (see Kamp 1981). Hence when you judge an assertion to be true, you have related either its initial linguistic representation, or a model based on that representation, to a model of the world. And, more important, you know that it is true: you are aware of having made a comparison, and such an awareness in turn depends on a model of your own performance.

Of course much of language goes beyond the perceptible. Some expressions refer directly to mental states, processes, and feelings; there is a vocabulary for referring to one's own internal milieu. Other expressions refer to abstract matters such as possibility, permissibility, and causation (Miller and Johnson-Laird 1976). These abstract concepts relate to scenarios, that is, models of a course of hypothetical or future events (Tversky and Kahneman 1973; Johnson-Laird 1983, pp. 410 et seq), and to conventions that regulate interactions between members of a society (Johnson-Laird 1983, pp. 415 et seq).

The case for discourse models has been advanced in formal semantics (Kamp 1981, Spencer-Smith 1987), in linguistics (Karttunen 1976, Reichgelt 1982, Shadbolt 1983, Fauconnier 1985), in artificial intelligence (Webber 1978, Wilks and Bien 1979), and in psycholinguistics (Stenning 1978, 1986, Johnson-Laird and Garnham 1980, Garnham 1981, 1987, Garnham and Oakhill 1989, van Dijk and Kintsch 1983, Glenberg, Meyer, and Lindem 1987). But there is a major problem for psychological theories. An assertion such as "A table is in front of the stove" can be true of an infinite number of different possible situations. Hence in formal semantics theorists postulate that an assertion has an infinite number of models, or "possible worlds," in which it would be true (see chapter 6). Granted that the mind has only a finite capacity, then an infinite number of models, as Partee (1979) has observed, cannot fit inside anyone's head. One solution to this problem is to assume that the initial linguistic representation of an assertion is used to construct just *one* model (Johnson-Laird 1983, ch. 11), but this model can serve as a representative and provisional sample from the infinite set of all possible models of the assertion. It can stand in for the correct model, presuming that the speaker has a specific state of affairs in mind, because it can be revised in the light of subsequent discourse. I will describe in the next section how the inferential procedure for such revisions could work.

In summary the theory of discourse models is based on three principal ideas:

1. A mental model represents the *reference* of a discourse, that is, the situation that the discourse describes.
2. The initial linguistic representation of a discourse, together with the machinery for constructing and revising discourse models from it, captures the *meaning* of the discourse, that is, the set of all possible situations that it could describe.
3. A discourse is judged to be true if there is at least one model of it that can be embedded in a model of the real world.

12.3 Reasoning and Mental Models

Three Theories of Reasoning

Inference is a systematic process of thought that leads from one set of propositions to another. Granted that the premises may be mentally represented in the form of a model, it is natural to ask how models might enter into inferential processes. The nature of these processes, however, is highly controversial in both artificial intelligence and cognitive psychology. The arguments in the two disciplines are largely independent of one another, but they run in parallel to a remarkable degree. There are the same three main points of view.

The first class of theories assumes that reasoning depends on formal rules of inference, like those of a logical calculus. The use of logic in artificial intelligence has been defended by Hayes (1977), and a variety of formal systems have been implemented (see, for example, Robinson 1979, Reiter 1973). The programming language Prolog is based on the same philosophy (Kowalski 1979). In psychology too there are many theories of reasoning that postulate a "mental logic" consisting of formal rules of inference (see, for example, Inhelder and Piaget 1958, Osherson 1975, Braine 1978, Rips 1983). Formal rules work in a purely syntactic way, and so they are blind to the content of a premise, depending solely on its so-called logical form—a notion that also has currency in linguistic theory (see Chomsky 1977, Hornstein 1986). Unlike such rules, however, people are highly sensitive to the content of premises when they make inferences. Such effects on the inferences of daily life have been independently discovered by workers in both artificial intelligence and cognitive psychology.

On the one hand psychologists have found that the difficulty of a deductive problem, and the nature of the responses to it, can be profoundly affected by its content (see, for example, Wason and Johnson-Laird 1972, Evans 1982). On the other hand there is a hiatus between what is valid in logic and in daily life. Logic, for example, warrants the inference from

If patients have cystitis, then they are given penicillin.

to the conclusion

If patients have cystitis and are allergic to penicillin, then they are given penicillin.

which is clearly an inference that runs counter to common sense. In logic, however, a conditional is treated as true whenever its antecedent is false (or its consequent is true), and so if the premise here is true, the conclusion must be true. Logic is indeed "monotonic" in that the validity of an inference is unaffected no matter what additional premises are added. Various attempts have been made to formulate "nonmonotonic" logics that tally with the inferences of daily life (see, for example, McDermott and Doyle 1980). The problem usually arises from the content of premises, however, not their logical form, and so it is unlikely to be patched up by a formal remedy (Davis 1980). Still worse, many verbal inferences in daily life are not derivable within a formal calculus at all. Some depend on the particular situation to which the premises refer (Johnson-Laird 1983, pp. 240, 261); some are plausible on the basis of general knowledge, but may be overruled by specific information to the contrary (see Minsky 1975, Schank and Abelson 1977); some derive from premises that can never be rendered sufficiently complete to ensure validity (Johnson-Laird 1987); and some are inductions (see chapter 13). The fact that so much reasoning is not deductive has led one former adherent of formal rules to abandon them (McDermott 1986).

The second class of theories directly recognize the importance of content. They postulate *content-specific* rules of inference. One origin of such theories lies in programming languages, such as PLANNER (Hewitt 1971), and production systems (Newell 1973) that enable general assertions to be expressed in the form of conditional rules, such as

If x is a dog, then x is an animal.

If someone asserts that Fido is a dog, then this rule will be triggered because the assertion matches its antecedent, and it will spring to life to make the further assertion that Fido is an animal. Another rule can be formulated so that given the goal of showing that Fido is an animal, it yields the subgoal of showing that Fido is a dog. Such rules are commonplace in the representation of the knowledge used in so-called expert systems (see chapter 14). These are computer programs that embody human expertise and that are designed to help their users to reach sensible decisions about such matters as medical diagnosis, chemical analysis, or where to drill for minerals. They rely on a large knowledge-base of rules that have been culled from interrogating human experts and on procedures that use these rules to make inferences about specific cases (see, for example, Buchanan and Feigenbaum 1978, Michie 1979, Davis and Lenat 1982, Feigenbaum and McCorduck 1984).

The idea of basing psychological theories of reasoning on content-

specific rules was discussed by Johnson-Laird and Wason (1977), and various sorts of such theories have been proposed (see, for example, Anderson 1983, E. R. Smith 1984, Cheng and Holyoak 1985, and Holland, Holyoak, Nisbett, and Thagard 1986). A related idea is that reasoning depends on the accumulation of specific examples within a connectionist framework, where the distinction between inference and recall is blurred (see chapter 4).

The main case for formal rules is that they explain how in principle people can reason about anything regardless of its content, including abstract and unfamiliar domains. The main case for content-specific rules is that they explain how the content of premises affects reasoning. But of course it is necessary to explain both these phenomena, as well as nondeductive inferences, and most relevant here is the third class of theories—those based on mental models. They do not employ rules of inference of any sort, either formal or content-specific, but assume instead that reasoning depends on the manipulation of mental models.

Such theories have been formulated for a variety of domains (see, for example, de Kleer and Brown 1981, Kahneman and Tversky 1982, Johnson-Laird 1983, ch. 5). Some rare individuals appear to have developed a conscious strategy that relies on the technique. As the late Richard Feynman explained (Feynman and Leighton 1985),

I had a scheme, which I still use today when somebody is explaining something that I'm trying to understand: I keep making up examples. For instance, the mathematicians would come in with a terrific theorem, and they're all excited. As they're telling me the conditions of the theorem, I construct something that fits all the conditions. You know, you have a set (one ball)—disjoint (two balls). Then the balls turn colors, grow hairs, or whatever, in my head as they put more conditions on. Finally they state the theorem, which is some dumb thing about the ball which isn't true for my hairy green ball thing, so I say, 'False!'

The same idea informs theories of reasoning based on mental models, which I illustrate first in terms of syllogistic inference.

Mental Models in Syllogistic Reasoning

Some syllogisms are so easy that even nine-year-old children can draw correct conclusions from them. Thus given the premises

None of the athletes is a beachcomber.

All the clerks are beachcombers.

most people correctly deduce the conclusion

None of the athletes is a clerk.

or its equally valid converse. Other syllogisms, however, are very much harder—so hard in fact that the majority of adults fail to draw a correct conclusion. Given the premises

None of the athletes is a beachcomber.

Some of the clerks are beachcombers.

most people make one of the following erroneous responses:

None of the athletes is a clerk.

None of the clerks is an athlete.

There's no valid conclusion.

Only a few draw the valid conclusion

Some of the clerks are not athletes.

The theory of mental models proposes that the reasoners' first task is to understand the premises and in particular to construct a model of them. Various proposals have been made about the form of such models. Some theorists assume that they are Euler circles, where sets are represented by circles that may or may not overlap (see Erickson 1974, Guyote and Sternberg 1981). One feature of this proposal is that there is no need to introduce a special symbol for negation. Other theorists concur that abstract notions such as negation should be encoded only in linguistic representations (Inder 1987, Jackendoff 1987). Euler circles, however, cannot represent assertions containing more than one quantifier, for example, "*Some* of the athletes know *all* the artists," and they can lead to combinatorial explosions because many premises call for several separate representations. Moreover there is a decisive objection to models that lack abstract elements. They cannot represent such assertions as "Ben knows that his presents weren't left by Santa Claus," because such "propositional attitudes" as the example shows often have a negative content. Another version of mental-model theory therefore makes the important assumption that any propositional attitude can itself be represented by a corresponding component within a mental model (Johnson-Laird 1983, ch. 15). To represent a negative proposition, for example, a special symbol for negation is directly introduced into a model. There is nothing improper about such a maneuver provided that the routines for evaluating the truth of models have an appropriate procedure for the symbol (see Kamp 1981; Johnson-Laird 1983, pp. 423-442). Indeed Venn diagrams are a traditional notation exploiting just such a device, that is, regions within three overlapping circles are shaded to represent the nonexistence of certain sets (see Newell 1981 for an algorithm for reasoning with Venn diagrams, and Polk and Newell 1988 for a defense of models containing propositional elements).

My colleagues and I have argued that the models used in reasoning are neither Euler circles nor Venn diagrams, because they are remote from the perceived structures of situations. We assume that the models are instead the discourse models discussed in the previous section. The premise

None of the athletes is a beachcomber.

is represented by a model containing an arbitrary number of tokens for athletes and an arbitrary number of tokens for beachcombers that are mentally tagged in some way to indicate that they are disjoint, for example,

athlete
athlete

beachcomber
beachcomber
beachcomber

where the barrier separates the two sets of tokens. (An alternative and perhaps equally plausible representation would tag each individual athlete as not a beachcomber, and vice versa.) The information from the premise

All the clerks are beachcombers.

can be directly added to the model

athlete
athlete

beachcomber = clerk
beachcomber = clerk
(beachcomber)

where the beachcomber who is not a clerk has been tagged with parentheses to represent the possible existence of such individuals.

The second stage of the process is the formulation of a putative conclusion: the model is scanned to determine what relation, if any, holds that is not explicitly stated in the premises. In the present case the procedure readily establishes the conclusion

None of the athletes is a clerk.

or its converse, depending on which direction the model is scanned.

The third stage consists of a search for a counterexample to the putative conclusion. An inference is valid if its conclusion cannot be false, given that its premises are true. Hence validity can be tested by searching for counterexamples—a procedure that has been exploited in logic (see, for example, Beth 1971). There is no need to manipulate the number of individuals for its own sake because it has no bearing on the conclusion. The model has only a finite number of tokens, and there are only a finite number of possible rearrangements of them. Because there is no way of establishing identities between athletes and clerks that does not also violate the premises, the present conclusion is valid. In fact even if no attempt to test it is made, it remains the correct answer.

Matters are very different in the case of the problem in which the second premise is

Some of the clerks are beachcombers.

If the information in this premise is added to the model in the following way:

athlete		
athlete		
<hr/>		
	beachcomber = clerk	
	beachcomber = clerk	
	(beachcomber)	(clerk)

then the model supports the erroneous conclusions

None of the athletes is a clerk.

None of the clerks is an athlete.

which the majority of subjects draw. A search for a counterexample can produce the following model:

athlete	=	(clerk)
athlete	=	(clerk)
<hr/>		
	beachcomber = clerk	
	beachcomber = clerk	
	(beachcomber)	(clerk)

Subjects who succeed in constructing it may nevertheless err by deciding that the premises fail to support any definite conclusion. It is necessary to scan the model from clerks to athletes to appreciate that all the possible models of the premises support the conclusion

Some (at least) of the clerks are not athletes.

Can we be certain that subjects follow such principles in reasoning? Higher cognitive processes tend to occur in many different ways, and some individuals may construct an initially misleading model and then revise it, whereas others may appreciate the existence of different possible models right from the start. One point is certain: without a training in logic, ordinary individuals do not have a simple standard procedure for dealing with syllogisms. Hence those premises that lead only to a single model are reliably easier than those that offer a choice of models. The precise number of distinct models that a subject constructs on any occasion is uncertain. My colleagues and I have developed a number of computer programs that model syllogistic reasoning and that differ on this point (see, for example, the two programs described in Johnson-Laird and Bara 1984). We can be sure, however, that many subjects do construct initially erroneous models—they draw invalid conclusions.

Similarly when subjects are given the chance to think again after they have had only a brief interval (10 s) in which to formulate a conclusion, they often change their minds (Johnson-Laird and Bara 1984). An unpublished experiment by Ruth Byrne and myself is relevant here: when subjects are given an unexpected recognition test of the conclusions that they have drawn to a series of different syllogisms, one common error is to select the conclusion predicted by the initial model in place of the subjects' actual (and correct) conclusion. This error is of course predicted if subjects draw an initial conclusion according to one model, which they then revise after they have constructed another model.

There is no doubt that people are able to search for counterexamples to conclusions, but the process is affected by the cognitive load of the task (see Oakhill and Johnson-Laird 1985). What is much harder to identify are the actual processes by which such models are constructed. Allen Newell in his William James Lectures at Harvard University in 1987 argued that mental models can be treated as state representations within a problem space. Thus when someone solves, say, the missionaries and cannibals problem, he or she applies a series of operations to transform a model of the initial state of affairs through a succession of models representing intermediate states until the goal is reached. This formulation is useful in characterizing the sequence of conscious states that an individual is aware of in solving a problem. One of the oddities of deductive reasoning, however, is that people have little conscious access to how they come up with a conclusion. Unless they are using visual images, they have no conscious access to the models themselves—an introspective deficiency that critics of mental models cite as contrary to the theory (Martin Braine, personal communication, 1988). Because the representation of individuals to whom anaphoric reference can be made is equally inaccessible—and there must be such a representation—I do not think that inaccessibility counts decisively against the theory. What remains true, however, is that little is known about the nature of the processes that generate counterexamples. A number of processes may occur in parallel and thus violate a direct analysis in terms of Newell's problem space, which assumes that only single operations lead from one state to the next within the space.

The crux is whether a valid conclusion calls for more than one model to be constructed. Whenever there is a choice of models, ordinary individuals are in serious danger of falling into error, because they lack a systematic inferential procedure.

Models and Other Forms of Reasoning

Reasoning on the basis of mental models depends, as shown in the previous subsection, on three semantic procedures:

1. The construction of a mental model of the state of affairs described in the premises, taking into account any relevant general and specific

knowledge. This procedure corresponds to the ordinary comprehension of discourse.

2. The formulation of a novel conclusion based on the model, unless of course a conclusion is already present for evaluation. This procedure corresponds to the description of a state of affairs with the proviso that the description should establish a relation not explicitly stated in the premises.

3. A search for alternative models that refute the putative conclusion. Only this search for counterexamples is peculiar to the process of inference. If there is no such model then the conclusion is valid. If there is such a model, then the reasoner must return to the second step and try to construct a new conclusion true in all the models so far constructed. If it is not clear whether there is such a model, then the conclusion can be accepted tentatively or expressed with some modal or probabilistic qualification (see Kahneman and Tversky 1982), but it may be subject to revision in the light of subsequent information.

The search for counterexamples solves the puzzle of how one model can stand in for an infinite number of different possible situations. Whenever an assertion is interpreted it may be necessary to make an arbitrary assumption to construct a single model. If a subsequent assertion is false in relation to this model, then it may be false because it conflicts with something depending on the arbitrary assumption. Hence an attempt can be made to revise the model so that it is consistent with the new assertion while still remaining an accurate model of the previous discourse. In this way earlier arbitrary assumptions can be corrected, or, should the revision be impossible, a genuine inconsistency can be detected between the latest assertion and the earlier discourse. This procedure is of course closely related to the search for counterexamples in deductive reasoning: there the aim is to render a currently true assertion false to check its validity; here the aim is to render a currently false assertion true to check its consistency. The same mechanism can be used to reason nonmonotonically. If an assumption is embodied in a model on the basis of default or prototypical information (see chapter 13), for example, that a dog has four legs, then a subsequent assertion may conflict with the model, for example, "my dog has three legs." In this case an attempt is made to revise the model. Such revisions can undo default assumptions (as well as arbitrary ones), but they cannot undo those conditions that are necessary to a concept.

The theory of inference based on mental models has been explored in a variety of domains. Indeed the study of three-term series problems, such as

Alice is taller than Bill.

Bill is taller than Charles.

Therefore Alice is taller than Charles.

led to one of the earliest versions of the theory. Various researchers, notably Huttenlocher (1968), proposed that reasoners form a mental arrangement of the individuals in the appropriate serial order. Other research revealed a more complicated picture. Thus Clark (1969) showed that a term such as "taller" is easier to understand than its converse "shorter," because the former is essentially neutral and affirmative in tone, whereas the latter is contrastive and negative in tone. A further complication is that subjects appear to develop different strategies to cope with the experimental task (compare Sternberg and Weil 1980, Egan and Grimes-Farrow 1982).

Spatial reasoning depends on more complex relations than three-term series problems. Consider, for example, the following inference:

The black ball is directly beyond the cue ball.

The green ball is on the right of the cue ball, and there is a red ball between them.

Hence if I move so that the red ball is between me and the black ball, then the cue ball will be to the left (of my line of sight).

Rules of inference for it would be complicated, and it is more likely to be drawn by manipulating a model of the spatial arrangement (see Johnson-Laird 1983, ch. 11, for the description of a spatial-reasoning program based on models).

One central point to make clear is that the logical properties of a term are emergent properties of its meaning rather than an explicit part of that meaning. For instance, the relation "greater than" has the logical property of transitivity, that is, any inference of the form

$x > y$ and $y > z$, therefore $x > z$

is valid. Yet there is no need to postulate a mental logic containing such a rule. Granted the concept of the successor of a number, for example, the successor of 2 is 3, the relation can be defined recursively (see Rogers 1967).

$x > y$ if x is the successor of y , or there is some number z such that x is the successor of z and z is greater than y .

It follows from this definition that, say, $4 > 2$, because there is a number, 3, such that 4 is its successor and $3 > 2$ (because 3 is the successor of 2). The principle of transitivity is plainly not part of the definition, but any model based on the meaning of the premises $a > b$ and $b > c$ yields the conclusion $a > c$.

The emergence of logical properties from meanings is a general principle. It applies to relations that hold between propositions, including simple relations such as "and" and "or," and more complicated conditional and causal relations. As in negation the key to the representation of these abstract relations is the introduction of appropriate procedures for interpreting special elements within models. A discourse may

describe an actual situation, a possible situation, or a hypothetical or fictitious situation. There must be therefore some way of representing the status of a discourse and of symbolizing it within a model. A conditional assertion, such as

If John is here, then Mary has left.

accordingly calls for a model in which the antecedent situation is represented as a possible state of affairs in which the consequent situation holds (see Johnson-Laird 1986).

Recent results suggest that conditional reasoning may call for a search for counterexamples rather than the manipulation of linguistic representations according to formal rules of inference. It is well known that people are inclined to commit certain fallacies. Thus given the premises

If it rains, she get wet.

It doesn't rain.

they conclude fallaciously *She does not get wet*. Why not postulate a corresponding fallacious formal rule? Because, say the defenders of mental logic, the inference can be suppressed by providing an appropriate additional premise suggesting that there are other ways in which to get wet, for example, if it snows, she gets wet (Rumain, Connell, and Braine 1983). Byrne (1989) has shown, however, that an additional premise can also suppress the use of the central rule of formal logic, modus ponens. Thus given the following sort of premises:

If it rains, she gets wet.

If she goes out, she gets wet.

It rains.

there is a striking suppression in the percentage of people who conclude *She gets wet*. The explanation, Byrne argues, is that people construct a model of the situation described by the premises, taking into account their general knowledge. They can readily find a counterexample to the putative conclusion that she gets wet, that is, they assume she does not go out. Thus they treat the original conditional, "If it rains, she gets wet," as at best an elliptical description of the state of affairs.

Any form of deductive reasoning in a finite domain can be based on semantic procedures for constructing, interpreting, and manipulating mental models. These procedures will obviously be sensitive to the content of premises. Likewise stress, emotionality, prejudice, or frank psychopathology may lead to a failure at any stage and particularly to a failure to confront counterexamples. On the one hand the theory allows for rationality in principle—the complete and correct performance of all the procedures; on the other hand it provides a natural explanation for errors.

12.4 Models as the Representation of Knowledge

Whenever you understand some phenomenon, such as how an electronic calculator operates, how a hydrogen bomb detonates, or how Gödel's theorem works, you have a mental representation of it that is like a working model. Expertise rests on such knowledge, and cognitive scientists have undertaken a massive investigation of systems that represent knowledge and retrieve it when it is needed. There have been studies of models in many domains, including the movement of objects (de Kleer 1977, de Kleer and Brown 1981, Forbus 1983), electrical circuits (Gentner and Gentner 1983, de Kleer 1985), propulsion systems (Williams, Hollan, and Stevens 1983), medical diagnosis (Meyer, Leventhal, and Gutmann 1985), navigation (see, for example, Oatley 1977, Hutchins 1983), and, above all, computing systems (see, for example, Du Boulay, O'Shea, and Monk 1981, Soloway, Ehrlich, Bonar, and Greenspan 1982).

Theorists have used the term "mental model" in these contexts to refer primarily to the *content* of a mental representation. But, although mental models may differ markedly in their content, there is no evidence to suggest that they differ in representational format or in the processes that construct and manipulate them. What is at issue is how such models develop as an individual progresses from novice to expert, and whether there is any pedagogical advantage in providing people with models of tasks they are trying to learn.

The Development of Expertise

In the study of intellectual development there has been a shift away from the traditional Piagetian emphasis on a change in structures and processes toward the view that what really changes is the content of knowledge (see, for example, Carey 1985, Keil 1979, 1989). Similarly as adults become more competent in a particular domain, they develop a richer model of that domain. As Young (1983) has argued, your model of an electronic calculator may represent merely a simple causal link from pushing certain keys to the execution of certain mathematical functions:

You type in an expression. The machine examines it, analyzes it, and calculates the answer according to the rules of arithmetic.

But to exploit the calculator fully, you need a rather richer model that maps specifications of particular arithmetic tasks onto specifications of particular sequences of actions—a so-called task-action model (Moran 1981, Young 1983, Green, Schiele, and Payne 1988). The model that you need to diagnose a malfunction will be still different again, perhaps incorporating the notion of binary adders, logic gates, and so on.

An important difference between the way a novice and an expert reason about a physical situation is that the novice's model represents

objects in the world and simulates processes that occur in real time, whereas a trained scientist can construct a model that represents highly abstract relations and properties, such as forces and momenta (Larkin 1983). The novice reasons qualitatively because appropriate quantitative reasoning calls for the more abstract model. A similar moral can be drawn from Tversky and Kahneman's work on judgments of probability. Naive individuals reason on the basis of mental simulations that call for the construction of models representing typical sequences of affairs (Kahneman and Tversky 1982); appropriate quantitative reasoning demands a more abstract model that embodies such factors as the a priori probabilities of events, the variances of distributions, and so on.

A model of a domain may be incomplete or inaccurate, and yet it can still be useful (Norman 1983; Johnson-Laird 1983, ch. 1)—just as a clock can be useful even though it is neither wholly accurate nor a complete representation of the earth's rotation. An erroneous model can obviously lead to erroneous conclusions and to certain persistent cognitive illusions. Thus, for example, many people believe that if a coin is rolled round the side of another fixed coin from top to bottom, it will end up inverted (see diSessa 1983). Many students believe that if an object is rotated on the end of a string, and the string snaps, then the trajectory of the object (ignoring gravity) is a spiral (see McCloskey, Caramazza, and Green 1980). Yet erroneous models are not always sources of error, and sometimes they may be a better guide than more sophisticated models. Kempton (1986) reports that there are two common models of thermostats. One model assumes that a thermostat acts as a feedback device that senses temperature and turns the furnace on or off to maintain a given temperature. The other, more primitive, model treats the thermostat as a valve that directly controls the furnace as the knob on a gas cooker. A "feedback" model can lead to a failure to set the thermostat low at night, if it implies that the fuel saved is balanced by the extra expense of reheating the house in the morning. This error is not made with the "valve" model because it implies that a lower setting always reduces the use of the furnace.

The source of error in a model may be mere ignorance, as in many misconceptions about force in physical systems. Sometimes, however, the error arises from a failure to envisage the situation properly or to hold in mind various possibilities, particularly in difficult deductive inferences. A nice example of the failure to envisage a situation has been described by Hinton (1979). The task is to imagine a cube balanced on one corner with the diametrically opposed corner vertically above it and then to indicate the locations of the other corners of the cube. Correct performance is rare without considerable previous experience with cubes. Many people consider that there are only four other corners that lie on the same horizontal plane.

Pedagogy, Analogy, and the Source of Models

One source of mental models is observation (aided by knowledge), another is other people's explanations, and still another is our ability to construct models for ourselves either from a set of basic components or from analogous models that we already possess. The whole of mathematics can be constructed from a small set of primitive ideas; any computational procedure can be constructed from a small set of building blocks; and most surprisingly explanations of the physical world, it seems, also rest on a foundation of basic ideas. Once we have a grasp of some of these ideas, a verbal explanation of a phenomenon enables us to construct a model of it. (Certain aspects of quantum physics elude such models; see Feynman 1985.)

A pedagogical precept, laid down by Wertheimer (1961), is that information should be presented to students in a way that enables them to cope with novel problems. This precept has also emerged from modern studies: an outline of the causal model of a machine, as opposed to its operating principles alone, enables novices to make inferences about novel problems (Halasz and Moran 1983, Kieras and Bovair 1984).

In situations where there is no teacher, people are reasonably adept at constructing causal models of their own. They already possess a rich knowledge of the variety of causal relations (Miller and Johnson-Laird 1976, sec. 6.3). They understand three important principles: first, in a deterministic domain all events have causes; second, causes precede their effects; and third, an action on an object is the likely cause of any change that occurs in it. These principles suffice, as Lewis (1986) has demonstrated in a computer program, for the construction of simple causal models of a physical system.

When a causal model fails to explain some phenomenon, a person is likely to search for a useful analogy, for example, the model of a thermostat as a valve. Gentner (1983, 1989) argues that the mere similarity of features between one domain and another cannot possibly account for the use of one as an analogy for the other. What has to be carried over are higher-order semantic relations, such as a causal framework. Thus in Rutherford's analogy between the solar system and the atom, the sun maps onto the nucleus of the atom, and the planets map onto the electrons. The properties of the sun, such as its heat, are dropped, but the higher-order relations are carried over. Hence the relation

the sun's attraction of the planets causes them to revolve around it
yields the inference

the nucleus's attraction of the electrons causes them to revolve
around it.

In fact the computer implementation of Gentner's theory establishes the best global mapping of systematic structure and then sets up the

mappings between the objects (Falkenhainer, Forbus, and Gentner 1986).

Holyoak and his colleagues have argued that the critical step in the mobilization of an analogy is the failure to solve a problem (Gick and Holyoak 1983, Holyoak 1985, Holyoak and Thagard 1989). The failure triggers an attempt to find an analogous problem to which the solution *is* known—a procedure that depends on finding existing links from the concepts active in the unsolved problem to those in another domain. Once sufficient links have been established, the actions contributing to the earlier solution can be transferred to the new problem. Analogical thinking is almost certainly not a unitary process, however, but, as Keane (1988) emphasizes, a set of different processes that depend on what an individual knows about a particular domain. In the case of profound scientific analogies, the search for links between one domain and another seems likely to be beyond the scope of any general algorithm.

12.5 A Critique of Mental Models

Theories of mental models have generated considerable interest, and inevitably they have also excited criticism, particularly from those committed to mental logic. One frequent objection is voiced in the form of a question: What exactly is a mental model? If the questioner requires a working definition, then a mental model can be defined as a representation of a body of knowledge—either long-term or short-term that meets the following conditions:

1. Its structure corresponds to the structure of the situation that it represents.
2. It can consist of elements corresponding only to perceptible entities, in which case it may be realized as an image, perceptual or imaginary. Alternatively it can contain elements corresponding to abstract notions; their significance depends crucially on the procedures for manipulating models.
3. Unlike other proposed forms of representation, it does not contain variables. Thus a *linguistic* representation of, say, *All artists are beekeepers* might take the form

For any x , if x is an artist, then x is a beekeeper.

In place of a variable, such as “ x ” in this expression, a model employs tokens representing a set of individuals.

Another major thrust against mental models is that they are an unnecessary explanatory concept. This criticism is similar to a familiar argument against imagery. Various authors, including Pylyshyn (1973, 1981) and Rips (1986), argue that everything can be represented by *propositional representations*, that is, structured strings of symbols in a

mental language. Any mental representation ultimately depends on neural events, and these, like the machine code of a computer, may well be computations on strings of symbols. But such symbols do not make explicit to consciousness the high level "relation structures" of which we are normally aware—the fact that the table is in front of the stove, and so on—and thus it seems that mental computations must have higher levels of organization. Indeed in another, narrower sense, the expression "propositional representation" is used to refer to high-level representations, namely, linguistic representations made up from symbols corresponding to lexical items in the language (see Kintsch 1974, Fodor 1975). In this case, as we have seen, there is a need to supplement them with mental models to account for reasoning and the comprehension of discourse.

The great danger for theories of representation is that they perpetrate the "symbolic fallacy" that meaning is merely a matter of relating one set of symbols to another. As Lewis (1972) said, to translate a sentence into a linguistic representation provides no more of an account of the conditions in which it is true than does a translation into Latin. One extreme reaction here is to say, in effect, so much the worse for truth conditions. Thus Rips (1986) wrote: "Cognitive psychology has to do without semantic notions like truth and reference that depend on the relationship between mental representations and the outside world." Alas, if we give up truth and reference, not much of mental life remains, and we cannot even account for the comprehension and verification of discourse. Contrary to Rips's methodological prescription many authors have urged that a major problem for cognitive science is to explain how symbols refer to the world (compare Hofstadter and Dennett 1981, Haugeland 1985, Russell 1987). The theory of mental models proposes a solution to this problem, though some commentators wrongly believe that it treats the interpretation of language as nothing more than the translation of utterances into models and neglects the question of how models are related to the world (see Oden 1987). As we have seen, however, the solution is that models of the world can also be constructed as a result of perception, internal experience, and social interaction. A discourse is deemed true if a model based on its linguistic representation can be embedded within such a model (Johnson-Laird 1983, pp. 247, 441).

A major problem with theories that invoke mental models as a representation of knowledge is their radical incompleteness. Indeed is it really possible to render overt our common sense about the everyday world, our knowledge of the meanings of words, and the competence underlying our expertise? A chorus of skeptics has argued to the contrary. Thus Dreyfus (1972), a persistent critic of artificial intelligence, says that the intractable problem in the simulation of human behavior is that "all alternatives must be made explicit." Even Winograd, an erstwhile advocate of artificial intelligence, has recently argued that

background knowledge cannot be represented as a set of explicit propositions (Winograd and Flores 1986). The overwhelming nature of knowledge may make it difficult—even impossible—for us to construct intelligent machines (other than by the original biological method). But it has no direct bearing on the feasibility of cognitive science, because a scientific account of how knowledge is acquired, retained, and used in interpreting the world does not necessarily call for a complete specification of all knowledge. Moreover, as Hayes (1979, 1985) and other like-minded theorists have argued, it may be possible to spell out intuitive knowledge in a completely explicit way. The enterprise is enormously time consuming, but if there are any insuperable barriers to it, they have yet to be discovered.

A more serious form of incompleteness concerns the theories of mental models themselves. Thus theories of vision do not yet give a full explanation of how retinal stimuli are mapped into rich dynamic models of the sort that the human perceptual system constructs. Theories of discourse account only for how fragments of language can be translated into models. Theories of reasoning have been advanced only for finite domains. And only a small number of expert systems have exploited model-based reasoning despite its apparent advantages for coping with nondeductive inferences. One reason for such deficiencies is the sheer difficulty of formulating theories, especially where the theorist is forced to analyze the truth conditions of expressions—a matter that can be finessed within semantic networks or other forms of linguistic representation.

12.6 Conclusion

In some domains theorists have emphasized the *content* of mental models. Thus studies of our everyday theories of the world have concentrated on what we know, and theorists have used a variety of representational formats—semantic networks, logical calculi, production systems—to represent such knowledge. In other domains, such as perception, discourse, and reasoning, theorists have emphasized the *structure* of models and how they are constructed by mental processes. Yet the theorists are talking about the same beast. Human beings create models of the world from both perceptual acquaintance with it and descriptions of it. These two forms of knowledge must be commensurable, otherwise we would never know what we were talking about or whether what we were saying was true. We retain such knowledge in long-term memory as the basis of our expertise in dealing with the world; and we can reason about any of our knowledge whether its source is perception, discourse, or memory.

In short, despite the diversity of theories, if we are to do justice to mental representations, the evidence suggests that those that we build from discourse are akin in structure to those that we build by other

means, and that reasoning exploits the same sort of models. Mental models are internal symbols, and so one question remains: What other sorts of symbols are there? Images, as I remarked, are a special sort of model—a two-dimensional representation that is projected from an underlying three-dimensional model. Hence the theory invokes a simple three-part inventory: linguistic representations, models, and procedures for manipulating them.

When Craik (1943) argued that people reason by carrying out thought experiments on internal models, the idea seemed dangerously heterodox. Now the range of phenomena that mental models are used to explain is growing rapidly. They include metacognition (Gilhooly 1986), consciousness and the self (Oatley 1981), intentional behavior and free will (Johnson-Laird 1988), and psychopathy (Power and Champion 1986). What remains as perhaps the major puzzle is how an entity can have recursive access to a model of its own performance (Weyrauch 1980; Johnson-Laird 1983, ch. 16; Smith 1984; Hayes-Roth, Garvey, Johnson, and Hewett 1987), and how it can exploit that knowledge in dealing with the world.

Note

I am grateful to my former colleague Alan Garnham, whose own ideas on mental models (see, for example, Garnham 1987) have had a major effect on my thinking. I thank Phil Barnard, Alan Garnham, Ed Smith, Mark Keane, and Michael Posner for many useful comments. Finally, I am grateful to Ruth Byrne for detailed criticisms of a previous draft and for other help in preparing this chapter.

References

- Anderson, J. R. 1983. *The Architecture of Cognition*. Cambridge, MA: Harvard University Press.
- Beth, E. W. 1971. *Aspects of Modern Logic*. Dordrecht: Reidel.
- Braine, M. D. S. 1978. On the relation between the natural logic of reasoning and standard logic. *Psychological Review* 85:1–21.
- Bransford, J. D., Barclay, J. R., and Franks, J. J. 1972. Sentence memory: A constructive versus interpretive approach. *Cognitive Psychology* 3:193–209.
- Buchanan, B. G., and Feigenbaum, E. A. 1978. DENDRAL and Meta-DENDRAL: Their applications dimension. *Artificial Intelligence* 11:5–24.
- Byrne, R. M. J. 1989. Suppressing valid inferences with conditionals. *Cognition* 31:61–83.
- Carey, S. 1985. *Conceptual Change in Childhood*. Cambridge, MA: MIT Press.
- Cheng, P. N., and Holyoak, K. J. 1985. Pragmatic reasoning schemas. *Cognitive Psychology* 17:391–416.

- Chomsky, N. 1977. *Essays on Form and Interpretation*. New York: North Holland.
- Clark, H. H. 1969. Linguistic processes in deductive reasoning. *Psychological Review* 79:387–404.
- Clark, H. H. 1977. Bridging. In P. N. Johnson-Laird and P. C. Wason, eds. *Thinking: Readings in Cognitive Science*. Cambridge, Engl.: Cambridge University Press.
- Craik, K. 1943. *The Nature of Explanation*. Cambridge, Engl.: Cambridge University Press.
- Davis M. 1980. The mathematics of non-monotonic reasoning. *Artificial Intelligence* 13:73–80.
- Davis, R., and Lenat, D. B. 1982. *Knowledge-based Systems in Artificial Intelligence*. New York: McGraw-Hill.
- de Kleer, J. 1977. Multiple representations of knowledge in a mechanics problem solver. *International Joint Conference on Artificial Intelligence* 5:299–304.
- de Kleer, J. 1985. How circuits work. In D. G. Bobrow, ed. *Qualitative Reasoning about Physical Systems*. Cambridge, MA: MIT Press.
- de Kleer, J., and Brown, J. S. 1981. Mental models of physical mechanisms and their acquisition. In J. Anderson, ed. *Cognitive Skills and their Acquisition*. Hillsdale, NJ: Erlbaum.
- diSessa, A. 1983. Phenomenology and the evolution of intuition. In D. Gentner and A. L. Stevens, eds. *Mental Models*. Hillsdale, NJ: Erlbaum, pp. 15–33.
- Dreyfus, H. L. 1972. *What Computers Can't Do: A Critique of Artificial Reason*. New York: Harper and Row.
- Du Boulay, B., O'Shea, T., and Monk, J. 1981. The black box inside the glass box: Presenting computer concepts to novices. *International Journal of Man-Machine Studies* 14:237–249.
- Egan, D. E., and Grimes-Farrow, D. D. 1982. Differences in mental representations spontaneously adopted for reasoning. *Memory and Cognition* 10:297–307.
- Ehrlich, K., and Johnson-Laird, P. N. 1982. Spatial descriptions and referential continuity. *Journal of Verbal Learning and Verbal Behavior* 21:296–306.
- Erickson, J. R. 1974. A set analysis theory of behaviour in formal syllogistic reasoning tasks. In R. Solso, ed. *Loyola Symposium on Cognition*. Vol. 2. Hillsdale, NJ: Erlbaum.
- Evans, J. St. B. T. 1982. *The Psychology of Deductive Reasoning*. London: Routledge and Kegan Paul.
- Falkenhainer, B., Forbus, K. D., and Gentner, D. 1986. The structure-mapping engine. University of Illinois, Technical Report No. UIUCDS-R-86-1275.
- Fauconnier, G. 1985. *Mental Spaces: Aspects of Meaning Construction in Natural Language*. Cambridge, MA: MIT Press.

- Feigenbaum, E. A., and McCorduck, P. 1984. *The Fifth Generation: Artificial Intelligence and Japan's Computer Challenge to the World*. London: Pan Books.
- Feynman, R. P. 1985. *QED: The Strange Theory of Light and Matter*. Princeton: Princeton University Press.
- Feynman, R. P., and Leighton, R. 1985. *"Surely You're Joking, Mr. Feynman!"* New York: W. W. Norton.
- Fodor, J. A. 1975. *The Language of Thought*. Hassocks, Sussex: Harvester Press.
- Forbus, K. D. 1983. Qualitative reasoning about space and motion. In D. Gentner and A. L. Stevens, eds. *Mental Models*. Hillsdale, NJ: Erlbaum, pp. 53–73.
- Garnham, A. 1981. Anaphoric reference to instances, instantiated and non-instantiated categories: A reading-time study. *British Journal of Psychology* 72:377–384.
- Garnham, A. 1987. *Mental Models as Representations of Discourse and Text*. Chichester: Ellis Horwood.
- Garnham, A., and Oakhill, J. V. 1987. Interpreting elliptical verb phrases. *Quarterly Journal of Experimental Psychology* 39A:611–627.
- Garnham, A., and Oakhill, J. V. 1989. The everyday use of anaphoric expressions: Implications for the 'Mental Models' theory of text comprehension. In N. E. Sharkey, ed., *Modelling Cognition: An Annual Review of Cognitive Science*. Vol. 2. Norwood, NJ: Ablex.
- Garnham, A., Oakhill, J. V., and Johnson-Laird, P. N. 1982. Referential continuity and the coherence of discourse. *Cognition* 1:29–46.
- Gentner D. 1983. Structure-mapping: A theoretical framework for analogy. *Cognitive Science* 7:155–170.
- Gentner, D. 1989. The mechanisms of analogical learning. In S. Vosniadou and A. Ortony, eds. *Similarity and Analogy in Reasoning and Learning*. Cambridge, Engl.: Cambridge University Press.
- Gentner, D., and Gentner, D. R. 1983. Flowing waters or teeming crowds: Mental models of electricity. In D. Gentner and A. L. Stevens, eds. *Mental Models*. Hillsdale, NJ: Erlbaum, pp. 99–129.
- Gick, M. L., and Holyoak, K. J. 1983. Schema induction and analogical transfer. *Cognitive Psychology* 15:1–38.
- Gilhooly, K. J. 1986. Mental modelling: A framework for the study of thinking. In J. Bishop, J. Lochhead and Perkins, D. N. eds. *Thinking: Progress in Research and Teaching*. Hillsdale, NJ: Erlbaum.
- Glenberg, A. M., Meyer, M., and Lindem, K. 1987. Mental models contribute to foregrounding during text comprehension. *Journal of Memory and Language* 26:69–83.
- Green, T. R. G., Schiele, F., and Payne, S. J. 1988. Formalisable models of user knowledge in human-computer interaction. In T. R. G. Green, G. C. van der Veer, J.-M. Hoc, and

- D. Murray, eds. *Working with Computers: Theory versus Outcome*. New York: Academic Press.
- Guyote, M. J., and Sternberg, R. J. 1981. A transitive-chain theory of syllogistic reasoning. *Cognitive Psychology* 13:461-525.
- Halasz, F. G., and Moran, T. P. 1983. Mental models and problem solving in using a calculator. *Proceedings of CHI '83: Human Factors in Computing Systems*. New York: Association for Computing Machinery.
- Haugeland, J. 1985. *Artificial Intelligence—The Very Idea*. Cambridge, MA: MIT Press.
- Hayes, P. J. 1977. In defense of logic. *Proceedings of the Fifth International Joint Conference on Artificial Intelligence*, pp. 559-565.
- Hayes, P. J. 1979. Naive physics 1—ontology for liquids. Memo, Centre pour les études Semantiques et Cognitives, Geneva. Reprinted in J. Hobbs and R. Moore, eds., 1985. *Formal Theories of the Commonsense World*. Hillsdale, NJ: Ablex, pp. 71-107.
- Hayes, P. J. 1985. The second naive physics manifesto. In J. Hobbs and R. Moore, eds. 1985. *Formal Theories of the Commonsense World*. Hillsdale, NJ: Ablex, pp. 1-20.
- Hayes-Roth, B., Garvey, A., Johnson, M. V., and Hewett, M. 1987. A modular and layered environment for reasoning about action. Knowledge Systems Laboratory Report No. KSL 86-38, Cognitive Science Department, Stanford University, Stanford, CA.
- Hewitt, C. 1971. PLANNER: A language for proving theorems in robots. *Fourth International Joint Conference on Artificial Intelligence*, pp. 115-121.
- Hinton, G. 1979. Some demonstrations of the effects of structural descriptions in mental imagery. *Cognitive Science* 3:231-250.
- Hofstadter, D. R., and Dennett, D. C., eds. 1981. *The Mind's I: Fantasies and Reflections on Self and Soul*. New York: Basic Books.
- Holland, J., Holyoak, K. J., Nisbett, R. E., and Thagard, P. 1986. *Induction: Processes of Inference, Learning, and Discovery*. Cambridge, MA: MIT Press. A Bradford Book.
- Holyoak, K. J. 1985. The pragmatics of analogical transfer. In G. H. Bower, ed. *The Psychology of Learning and Motivation: Advances in Research and Theory*. Vol. 19. New York: Academic Press.
- Holyoak, K. J. and Thagard, P. R. 1989. Rule-based spreading activation and analogical transfer. In S. Vosniadou and A. Ortony, eds. *Similarity and Analogy in Reasoning and Learning*. Cambridge, Engl.: Cambridge University Press.
- Hornstein, N. 1986. *Logic as Grammar*. Cambridge, MA: MIT Press.
- Hutchins, E. 1983. Understanding Micronesian navigation. In D. Gentner and A. L. Stevens, eds. *Mental Models*. Hillsdale, NJ: Erlbaum, pp. 191-225.
- Huttenlocher, J. 1968. Constructing spatial images: A strategy in reasoning. *Psychological Review* 75:550-560.

- Inder, R. 1987. The computer simulation of syllogism solving using restricted mental models. Ph.D. thesis, Cognitive Studies, Edinburgh University.
- Inhelder, B., and Piaget, J. 1958. *The Growth of Logical Thinking from Childhood to Adolescence*. New York: Basic Books.
- Jackendoff, R. 1987. On beyond Zebra: The relation of linguistic and visual information. *Cognition* 26:89–114.
- Johnson-Laird, P. N. 1983. *Mental Models: Towards a Cognitive Science of Language, Inference, and Consciousness*. Cambridge, MA: Harvard University Press. Cambridge, Engl.: Cambridge University Press.
- Johnson-Laird, P. N. 1986. Conditionals and mental models. In E. C. Traugott, A. ter Meulen, J. S. Reilly and C. A. Ferguson, eds. *On Conditionals*. Cambridge, Engl.: Cambridge University Press.
- Johnson-Laird, P. N. 1987. Reasoning, imagining, and creating. *Bulletin of the British Psychological Society* 40:121–129.
- Johnson-Laird, P. N. 1988. Freedom and constraint in creativity. In Sternberg, R. J., ed. *The Nature of Creativity*. New York: Cambridge University Press.
- Johnson-Laird, P. N., and Bara, B. G. 1984. Syllogistic inference. *Cognition* 16:1–61.
- Johnson-Laird, P. N., and Garnham, A. 1980. Descriptions and discourse models. *Linguistics and Philosophy* 3:371–393.
- Johnson-Laird, P. N., Herrmann, D. J., and Chaffin, R. 1984. Only connections: A Critique of semantic networks. *Psychological Bulletin* 96:292–315.
- Johnson-Laird, P. N., and Wason, P. C., eds. 1977. *Thinking: Readings in Cognitive Science*. Cambridge, Engl.: Cambridge University Press.
- Kahneman, D., and Tversky, A. 1982. The simulation heuristic. In D. Kahneman, P. Slovic, and A. Tversky, eds. 1982. *Judgement Under Uncertainty: Heuristics and Biases*. Cambridge, Engl.: Cambridge University Press, pp. 201–208.
- Kamp, J. A. W. 1981. A theory of truth and semantic representation. In J. A. G. Groenendijk, T. Janssen and M. Stokhof, eds. *Formal Methods in the Study of Language*. Amsterdam: Mathematical Center Tracts, pp. 277–322.
- Karttunen, L. 1976. Discourse referents. In J. McCawley, ed. *Syntax and Semantics, Vol. 7: Notes from the Linguistic Underground*. New York: Academic Press.
- Keane, M. 1988. *Analogical Problem Solving*. Chichester, Engl.: Ellis Horwood.
- Keil, F. 1979. *Semantic and Conceptual Development: Ontological Perspective*. Cambridge, MA: Harvard University Press.
- Keil, F. 1989. *Concepts, Word Meanings, and Cognitive Development*. Cambridge, MA: Harvard University Press.
- Kempton, W. 1986. Two theories of home heat control. *Cognitive Science* 10:75–90.

- Kieras, D. E., and Bovair, S. 1984. The role of a mental model in learning to operate a device. *Cognitive Science* 8:255–273.
- Kintsch, W. 1974. *The Representation of Meaning in Memory*. Hillsdale, NJ: Erlbaum.
- Kintsch, W., and van Dijk, T. A. 1978. Towards a model of text comprehension and reproduction. *Psychological Review* 85:363–394.
- Kowalski, R. A. 1979. *Logic for Problem Solving*. Amsterdam: Elsevier North-Holland.
- Larkin, J. H. 1983. The role of problem representation in physics. In D. Gentner and A. L. Stevens, eds. *Mental Models*. Hillsdale, NJ: Erlbaum, pp. 75–98.
- Lewis, C. 1986. A model of mental model construction. *Proceedings of CHI '86 Conference on Human Factors in Computer Systems*. New York: Association for Computing Machinery.
- Lewis, D. K. 1972. General semantics. In D. Davidson and G. Harman, eds. *Semantics of Natural Language*. Dordrecht: Reidel.
- McCloskey, M., Caramazza, A., Green, G. 1980. Curvilinear motion in the absence of external forces: Naive beliefs about the motion of objects. *Science* 210:1139–1141.
- McDermott, D. 1986. A critique of pure reason. Mimeo, Artificial Intelligence, Yale University, New Haven, CT.
- McDermott, D., and Doyle, J. 1980. Non-monotonic logic I. *Artificial Intelligence* 13:41–72.
- Mani, K., and Johnson-Laird, P. N. 1982. The mental representation of spatial descriptions. *Memory and Cognition* 10:181–187.
- Marr, D. 1982. *Vision: A Computational Investigation in the Human Representation of Visual Information*. San Francisco: Freeman.
- Meyer, D., Leventhal, H., and Gutmann, M. 1985. Common-sense models of illness: The example of hypertension. *Health Psychology* 4:115–135.
- Michie, D., ed. 1979. *Expert Systems in the Micro-Electronic Age*. Edinburgh: Edinburgh University Press.
- Miller, G. A. 1972. English verbs of motion: a case study in semantics and lexical memory. In A. W. Melton and E. Martin, eds. *Coding Processes in Human Memory*. Washington, DC: Winston.
- Miller, G. A., Galanter, E., and Pribram, K. 1960. *Plans and the Structure of Behavior*. New York: Holt, Rinehart, and Winston.
- Miller, G. A., and Johnson-Laird, P. N. 1976. *Language and Perception*. Cambridge, MA: Harvard University Press. Cambridge, Engl.: Cambridge University Press.
- Minsky, M. 1975. A framework for representing knowledge. In P. H. Winston, ed. *The Psychology of Computer Vision*. New York: McGraw-Hill.

- Moran, T. P. 1981. The Command Language Grammar: A representation for the user interface of interactive computer systems. *International Journal of Man-Machine Studies* 15:5–30.
- Newell, A. 1973. Production systems: Models of control structures. In W. G. Chase, ed. *Visual Information Processing*. New York: Academic Press.
- Newell, A. 1981. Reasoning, problem solving and decision processes: The problem space as a fundamental category. In R. Nickerson, ed. *Attention and Performance*. Vol. 8. Hillsdale, NJ: Erlbaum.
- Newell, A., and Simon, H. A. 1976. Computer science as empirical inquiry: Symbols and search. *Communications of the Association for Computing Machinery* 19:113–126.
- Norman, D. A. 1983. Some observations on mental models. In D. Gentner and A. L. Stevens, eds. *Mental Models*. Hillsdale, NJ: Erlbaum, pp. 7–14.
- Oakhill, J. V., and Garnham, A. 1985. Referential continuity, transitivity, and the retention of spatial descriptions. *Language and Cognitive Processes* 1:149–162.
- Oakhill, J. V., and Johnson-Laird, P. N. 1985. Rationality, memory, and the search for counterexamples. *Cognition* 20:79–94.
- Oatley, K. G. 1977. Inference, navigation, and cognitive maps. In P. N. Johnson-Laird and P. C. Wason, eds. *Thinking: Readings in Cognitive Science*. Cambridge, Engl.: Cambridge University Press, pp. 537–547.
- Oatley, K. G. 1981. Representing ourselves: Mental schemata, computational metaphors, and the nature of consciousness. In G. Underwood and R. Stevens, eds. *Aspects of Consciousness, vol. 2: Structural Issues*. New York: Academic Press.
- Oden, G. C. 1987. Concept, knowledge, and thought. *Annual Review of Psychology* 38:203–227.
- Osherson, D. N. 1975. Logic and models of logical thinking. In R. J. Falmagne, ed. *Reasoning: Representation and Process in Children and Adults*. Hillsdale, NJ: Erlbaum.
- Partee, B. H. 1979. Semantics—mathematics or psychology? In R. Baurle, U. Egli, and A. von Stechow, eds. *Semantics from Different Points of View*. Berlin: Springer.
- Polk, T., and Newell, A. 1988. Modeling human syllogistic reasoning in Soar. *Proceedings of the Tenth Annual Conference of the Cognitive Science Society*. Hillsdale, NJ: Erlbaum, pp. 181–187.
- Pylyshyn, Z. 1973. What the mind's eye tells the mind's brain: A critique of mental imagery. *Psychological Bulletin* 80:1–24.
- Pylyshyn, Z. 1981. The imagery debate: Analogue media versus tacit knowledge. In N. Block, ed. *Imagery*. Cambridge, MA: MIT Press.
- Power, M. J., and Champion, L. A. 1986. Cognitive approaches to depression: A theoretical critique. *British Journal of Clinical Psychology* 25:201–212.

- Reichardt, W. E., and Poggio, T. 1981. Visual control of flight in flies. In W. E. Reichardt and T. Poggio, eds. *Theoretical Approaches in Neurobiology*. Cambridge, MA: MIT Press.
- Reichgelt, H. 1982. Mental models and discourse. *Journal of Semantics* 1:371–386.
- Reiter, R. 1973. A semantically guided deductive system for automatic theorem-proving. *Third International Joint Conference on Artificial Intelligence*, pp. 41–46.
- Rips, L. J. 1983. Cognitive processes in propositional reasoning. *Psychological Review* 90:38–71.
- Rips, L. J. 1986. Mental muddles. In M. Brand and R. M. Harnish, eds. *Problems in the Representation of Knowledge and Belief*. Tucson, AZ: University of Arizona Press.
- Robinson, J. A. 1979. *Logic: Form and Function, The Mechanization of Deductive Reasoning*. Edinburgh: Edinburgh University Press.
- Rogers, H. 1967. *Theory of Recursive Functions and Effective Computability*. New York: McGraw-Hill.
- Rumain, B., Connell, J., and Braine, M. D. S. 1983. Conversational comprehension processes are responsible for reasoning fallacies in children as well as adults: IF is not the biconditional. *Developmental Psychology* 19:471–481.
- Russell, J. 1987. Rule-following, mental models, and the developmental view. In M. Chapman and R. A. Dixon, eds. *Meaning and the Growth of Understanding: Wittgenstein's Significance for Developmental Psychology*. New York: Springer.
- Sag, I., and Hankamer, J. 1984. Toward a theory of anaphoric processing. *Linguistics and Philosophy* 7:325–345.
- Sanford, A., and Garrod, S. 1981. *Understanding Written Language: Explorations of Comprehension Beyond the Sentence*. Chichester, Engl.: Wiley.
- Schank, R. C., and Abelson, R. P. 1977. *Scripts, Plans, Goals and Understanding*. Hillsdale, NJ: Erlbaum.
- Shadbolt, N. 1983. Processing reference. *Journal of Semantics* 2:63–98.
- Smith, B. C. 1984. Reflection and semantics in LISP. Conference Record of the Eleventh Annual Association for Computing Machinery Symposium on Principles of Programming Languages, Salt Lake City, Utah, pp. 23–35.
- Smith, E. R. 1984. Models of social inference processes. *Psychological Review* 91:392–413.
- Soloway, E., Ehrlich, K., Bonar, J., and Greenspan, J. 1982. What do novices know about programming? In A. Badre and B. Schneiderman, eds. *Directions in Human-Computer Interactions*. Norwood, NJ: Ablex, pp. 27–54.
- Spencer-Smith, R. 1987. Survey: Semantics and discourse representation. *Mind and Language* 2:1–26.
- Stenning, K. 1978. Anaphora as an approach to pragmatics. In M. Halle, J. Bresnan and G. A. Miller, eds. *Linguistic Theory and Psychological Reality*. Cambridge, MA: MIT Press.

- Stenning, K. 1986. On making models: A study of constructive memory. In T. Myers, K. Brown and B. McGonigle, eds. *Reasoning and Discourse Processes*. London: Academic Press.
- Sternberg, R. J., and Weil, E. M. 1980. An aptitude x strategy interaction in linear syllogistic reasoning. *Journal of Educational Psychology* 72:226-239.
- Tversky, A., and Kahneman, D. 1973. Availability: A heuristic for judging frequency and probability. *Cognitive Psychology* 4:207-232. Reprinted in Kahneman, D., Slovic, P., and Tversky, A., eds. 1982. *Judgement under Uncertainty: Heuristics and Biases*. Cambridge, Engl.: Cambridge University Press.
- van Dijk, T. A., and Kintsch, W. 1983. *Strategies of Discourse Comprehension*. New York: Academic Press.
- Wason, P. C., and Johnson-Laird, P. N. 1972. *Psychology of Reasoning: Structure and Content*. Cambridge, MA: Harvard University Press. London: Batsford.
- Webber, B. L. 1978. Description formation and discourse model synthesis. In D. L. Waltz, ed. *Theoretical Issues in Natural Language Processing*, 2. New York: Association for Computing Machinery.
- Wertheimer, M. 1961. *Productive Thinking*. London: Tavistock.
- Weyrauch, R. W. 1980. Prolegomena to a theory of mechanized formal reasoning. *Artificial Intelligence* 13:133-170.
- Wilks, Y., and Bien, J. S. 1979. Speech acts and multiple environments. *Sixth International Joint Conference on Artificial Intelligence*, pp. 451-455.
- Williams, D., Hollan, J. D., and Stevens, A. L. 1983. Human reasoning about a simple physical system. In D. Gentner and A. L. Stevens, eds. *Mental Models*. Hillsdale, NJ: Erlbaum, pp. 131-153.
- Winograd, T., and Flores, F. 1986. *Understanding Computers and Cognition: A New Foundation for Design*. Norwood, NJ: Ablex.
- Wittgenstein, L. 1922. *Tractatus Logico-Philosophicus*. London: Routledge and Kegan Paul.
- Young, R. 1983. Surrogates and mappings: Two kinds of conceptual models for interactive devices. In D. Gentner and A. L. Stevens, eds. *Mental Models*. Hillsdale, NJ: Erlbaum, pp. 35-52.