Chapter 11

Space to Think

Philip N. Johnson-Laird

11.1 Introduction

Perception is the transformation of local information at the sensorium into a mental model of the world at a distance, thinking is the manipulation of such models, and action is guided by its results. This account of human cognition goes back to the remarkable Scottish psychologist, Kenneth Craik (1943), and it has provided both a program of research for the study of human cognition and a central component of the theory of mental representations. Thus the final stage of visual perception, according to Marr (1982), delivers a three-dimensional model of the world, which the visual system has inferred from the pattern of light intensities falling on the retinas. Mental models likewise underlie one account of verbal comprehension: to understand discourse is, on this account, to construct a mental model of the situation that it describes (see, for example, Johnson-Laird 1983; Garnham 1987). The author and his colleagues have developed this account into a theory of reasoning—both inductive and deductive—in which thinkers reason by manipulating models of the world (see, for example, Johnson-Laird and Byrne 1991).

The idea of mental models as the basis for deductive thinking has its origins in the following idea:

> Consider the inference
>
> The box is on the right of the chair,
> The ball is between the box and the chair,
> Therefore, the ball is on the right of the chair.
>
> The most likely way in which such an inference is made involves setting up an internal representation of the scene depicted by the premises. This representation may be a vivid image or a fleeting abstract delineation—its substance is of no concern. The crucial point is that its formal properties mirror the spatial relations of the scene so that the conclusion can be read off in almost as direct a fashion as from an actual array of objects. It may be objected, however, that such a depiction of the premises is unnecessary, that the inference can be made

> by an appeal to general principles, or rules of inference, which indicate that items related by *between* must be collinear, etc. However, this view—that relational terms are tagged according to the inference schema they permit—founders on more complex inferences. An inference of the following sort, for instance, seems to be far too complicated to be handled without constructing an internal representation of the scene:
>
> 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.
> Therefore, if I move so that the red ball is between me and the black ball, then the cue ball is on my left.
>
> Even if it is possible to frame inference schema that permit such inferences to be made without the construction of an internal representation, it is most unlikely that this approach is actually adopted in making the inference. (Johnson-Laird 1975, 12–13)

This passage captures the essence of the model theory of deduction, but the intuition that spatial inferences are made by imagining spatial scenes turned out *not* to be shared by all investigators.

Twenty years have passed since the argument above was first formulated, and so the aim of this chapter is, in essence, to bring the story up to date. It contrasts the model theory with an account based on formal rules of inference, and it presents evidence that spatial reasoning is indeed based on models. It then argues that spatial models may underlie other sorts of thinking—even thinking that is not about spatial relations. It presents some new results showing that individuals often reason about temporal relations by constructing quasi-spatial models. Finally, it demonstrates that one secret in using diagrams as an aid to thinking is that their spatial representations should make alternative possibilities explicit.

11.2 Propositional Representations and Mental Models

What does one mean by a mental model? The essence of the answer is that its structure corresponds to the structure of what it represents. A mental model is accordingly similar in structure to a physical model of the situation, for example, a biochemist's model of a molecule, or an architect's model of a house. The parts of the model correspond to the relevant parts of the situation, and the structural relations between the parts of the model are analogous to the structural relations in the world. Hence, individual entities in the situation will be represented as individuals in the model, their properties will be represented by their properties in the model, and the relations among them will be represented by relations among them in the model. Mental models are partial in that they represent only certain aspects of the situation, and they thus correspond to many possible states of affairs, that is, there is a many-to-one mapping from situations in the world to a model. Images, too, have these properties, but models need not be visualizable, and unlike images, they may represent several distinct sets of possibilities. These abstract characterizations are hard to follow, but they can be clarified by contrasting mental models with so-called propositional representations.

To illustrate a propositional representation, consider the assertion:

A triangle is on the right of a circle.

Its propositional representation relies on some sort of predicate argument structure, such as the following expression in the predicate calculus:

$(\exists x)(\exists y)(\text{Triangle}(x) \mathbin{\&} \text{Circle}(y) \mathbin{\&} \text{Right-of}(x, y))$,

where $\exists$ denotes the existential quantifier "for some" and the variables range over individuals in the domain of discourse, i.e. the situation that is under description. The expression can accordingly be paraphrased in "Loglish"—a hybrid language spoken only by logicians—as follows:

For some x and for some y, such that x is a triangle and y is a circle, x is on the right of y.

The information in the further assertion

The circle is on the right of a line

can be integrated to form the following expression representing both assertions:

$(\exists x)(\exists y)(\exists z)(\text{Triangle}(x) \mathbin{\&} \text{Circle}(y) \mathbin{\&} \text{Line}(z) \mathbin{\&} \text{Right-of}(x, y) \mathbin{\&} \text{Right-of}(y, z))$.

A salient feature of this representation is that its structure does *not* correspond to the structure of what it represents. The key component of the propositional representation is

$\text{Right-of}(x, y) \mathbin{\&} \text{Right-of}(y, z)$,

in which there are four tokens representing variables. In contrast, the situation itself has three entities in a particular spatial relation. Hence, a mental model of the situation must have the same structure, which is depicted in the following diagram:

| ○ △

where the horizontal dimension corresponds to the left-to-right dimension in the situation. In what follows, such diagrams are supposed to depict mental models, and will often be referred to as though they were mental models. Each token in the present mental model has a property corresponding to the shape of the entity it represents, and the three tokens are in a spatial relation corresponding to the relation between the three entities in the situation described by the assertions. In the case of such a spatial model, a critical feature is that elements in the model can be accessed and updated in terms of parameters corresponding to axes.

The process of inference for propositional representations calls for a system based on rules, and psychologists have proposed such systems for spatial inference based on formal rules of inference (see, for example, Hagert 1984; Ohlsson 1984). Hence, in order to infer from the premises above the valid conclusion

A triangle is on the right of a line,

it is necessary to rely on a statement of the transitivity of "on the right of":

$(\forall x)(\forall y)(\forall z)((\text{Right-of}(x,y) \mathbin{\&} \text{Right-of}(y,z)) \rightarrow \text{Right-of}(x,z)),$

where ∀ denotes the universal quantifier "for any" and → denotes material implication ("if ..., then ..."). With this additional premise (a so-called meaning postulate) and a set of rules of inference for the predicate calculus, the conclusion can be derived in the following chain of inferential steps.

The premises are

(1) $(\exists x)(\exists y)(\text{Triangle}(x) \mathbin{\&} \text{Circle}(y) \mathbin{\&} \text{Right-of}(x,y))$

(2) $(\exists y)(\exists z)(\text{Circle}(y) \mathbin{\&} \text{Line}(z) \mathbin{\&} \text{Right-of}(y,z))$

(3) $(\forall x)(\forall y)(\forall z)((\text{Right-of}(x,y) \mathbin{\&} \text{Right-of}(y,z)) \rightarrow \text{Right-of}(x,z))$

The proof calls for the appropriate instantiations of the quantified variables, that is, one replaces the quantified variables by constants denoting particular entities:

(4) $(\exists y)(\text{Triangle}(a) \mathbin{\&} \text{Circle}(y) \mathbin{\&} \text{Right-of}(a,y))$ [from (1)]

(5) $(\text{Triangle}(a) \mathbin{\&} \text{Circle}(b) \mathbin{\&} \text{Right-of}(a,b))$ [from (4)]

(6, 7) $(\text{Circle}(b) \mathbin{\&} \text{Line}(c) \mathbin{\&} \text{Right-of}(b,c))$ [from (2)]

There are constraints on the process of instantiating variables that are existentially quantified, but universal quantifiers range over all entities in the domain, and so the meaning postulate can be freely instantiated as follows:

(8–10) $((\text{Right-of}(a,b) \mathbin{\&} \text{Right-of}(b,c)) \rightarrow \text{Right-of}(a,c))$ [from (3)]

The next steps use formal rules of inference for the connectives. A rule for conjunction stipulates that given a premise of the form $(A \mathbin{\&} B)$, where A and B can denote compound assertions of any degree of complexity, one can derive the conclusion B. Hence one can detach part of line 5 as follows:

(11) $\text{Right-of}(a,b)$ [from (5)]

and part of line 7 as follows:

(12) Right-of(b, c) [from (7)]

Another rule allows any two assertions in separate lines to be conjoined, that is, given premises of the form A, B, one can derive the conclusion (A & B). This rule allows a conjunction to be formed from the previous two lines in the derivation:

(13) (Right-of(a, b) & Right-of(b, c)) [from (11), (12)]

This assertion matches the antecedent of line 10, and a rule known as "modus ponens" stipulates that given any premises of the form ($A \rightarrow B$), A, one can derive the conclusion B. The next step of the derivation proceeds accordingly:

(14) Right-of(a, c) [from (10, (13)]

The rules for conjunction allow the detachment of propositions from previous lines and their assembly in the following conclusion:

(15–18) ((Triangle(a) & Line(c)) & Right-of(a, c)) [from (5), (7), (14)]

Finally, this propositional representation can be translated back into English:

Therefore, the triangle is on the right of the line.

The process of inference for models is different. The theory relies on the following simple idea. A valid deduction, by definition, is one in which the conclusion must be true if the premises are true. Hence what is needed is a model-based method to test for this condition. Assertions can be true in indefinitely many different situations, and so it is out of the question to test that a conclusion holds true in all of them. But testing can be done in certain domains precisely because a mental model can stand for indefinitely many situations. Here, in principle, is how it is done for spatial inferences. Consider, again, the example above:

A triangle is on the right of a circle.
The circle is on the right of a line.

The assertions say nothing about the actual distances between the objects. Instead of trying to envisage all the different possible situations that satisfy these premises, a mental model leaves open the details and captures only the structure that all the different situations have in common:

| ○ △

where the left-to-right axis corresponds to the left-right axis in space, but the distances between the tokens have no significance. This model represents only the spatial sequence of the objects, and it is the only possible model of the premises, that is, no other model corresponding to a different left-to-right sequence of the three objects satisfies the premises. Now consider the further assertion:

The triangle is on the right of the line.

It is true in the model, and, because there are no other models of the premises, it *must* be true given that the premises are true. The deduction is valid, and because reasoners can determine that there are no other possible models of the premises, they can not only make this deduction but also *know* that it is valid (see Barwise 1993).

The same principles allow us to determine that an inference is invalid. Given, say, the inference

A triangle is on the right of a circle,
A line is on the right of the circle,
Therefore, the triangle is on the right of the line,

the first premise yields the model

○ △

but now when we try to add the information from the second premise, the relation between the triangle and the line is uncertain. One way to respond to such an indeterminacy is to build separate models for each possibility:

○ | △   ○ △ |

ignoring the possibility that the triangle and the line might be, say, one on top of the other. The first of these models shows that the putative conclusion is possible, but the second model is a counterexample to it. It follows that the triangle *may* be on the right of the line, but it does not follow that the triangle *must* be on the right of the line.

Does the model theory abandon the idea of propositional representations? Not at all. It turns out to be essential to have a representation of the meaning of an assertion independent of its particular realization in a model. The theory accordingly assumes that the first step in recovering the meaning of a premise is the construction of its propositional representation—a representation of the truth conditions of the premise. This representation is then used to update the set of models of the premises.

The use of mental models in reasoning has two considerable advantages over the use of formal rules. The first advantage is that it yields a decision procedure—at least for domains such as spatial reasoning that can have one, because the predicate calculus is provably without any possible decision procedure. An inference is valid if its conclusion holds in all the possible models of the premises, and it is invalid if it fails to hold in at least one of the possible models of the problems. Granted that problems remain within the capacity of working memory, then it is a simple matter to decide whether or not an inference is valid. One examines the models of the premises, and a conclusion is valid if, and only if, it is true in all of them. The situation is very different in the case of formal rules. They have no decision procedure. Quine (1974, 75) commented on this point in contrasting a semantic decision procedure for the propositional calculus (akin in some ways to the mental model account of that domain) and an approach based on formal rules. Of the use of formal rules, he wrote: "It is inferior in that it affords no general way of reaching a verdict of invalidity; failure to discover a proof for a schema can mean either invalidity or mere bad luck." The same problem, as Barwise (1993) has pointed out, haunts psychological theories based on formal rules. The search space of possible derivations is vast, and thus such theories have to assume that reasoners explore it for a certain amount of time and then give up. Barwise remarks: "The 'search till you're exhausted' strategy gives one at best an educated, correct guess that something does not follow" (337). Models allow individuals to *know* that there is no valid conclusion.

The second advantage of mental models is that they extend naturally to inductive inferences and to the informal arguments of daily life to which it is so hard, if not impossible, to apply formal rules of inference (see, for example, Toulmin 1958). Such inferences and arguments nevertheless differ in their strength (Osherson, Smith, and Shafir 1986). The model theory implies that the strength of an inference—any inference—depends on the believability of its premises and on the proportion of models of the premises in which the conclusion is true (Johnson-Laird 1994). Hence the model theory provides a unified account of inference:

- If the conclusion holds in all possible models of the premises, it is *necessary* given the premises, that is, deductively valid.
- If it holds in most of the models of the premises, then it is *probable.*
- If it holds in some model of the premises, then it is *possible.*
- If it holds in only a few models of the premises, then it is *improbable.*
- If it holds in none of the models of the premises, then it is *impossible*, that is, inconsistent with the premises.

The theory forms a bridge between models and the heuristic approach to judgments of probability based on scenarios (see, for example, Tversky and Kahneman 1973). As the number of indeterminacies in premises increases, there is an exponential growth in the number of possible models. Hence the procedure is intractable for all but small numbers of indeterminacies. However, once individuals have constructed a model in which a highly believable conclusion holds, they tend not to search for alternative models that refute the conclusion. The theory according provides a mechanism for *inferential satisficing* (cf. Simon 1959). This mechanism accounts for the common failure to consider alternative lines of argument—a failure shown by studies of inference, both deductive (e.g., Johnson-Laird and Byrne 1991) and informal (e.g., Perkins, Allen, and Hafner 1983; Kuhn 1991), and by many real-life disasters, for example, the operators at Three Mile Island inferred that a relief valve was leaking and overlooked the possibility that it was stuck open.

11.3 Algorithm for Spatial Reasoning Based on Mental Models

The machinery required for reasoning by model calls, not for formal rules of inference, but procedures for constructing models, formulating conclusions true in models, and testing whether conclusions are true in models. The present author has implemented computer programs that make inferences using such an algorithm for syllogisms, sentential connectives, doubly quantified assertions, and several other domains including spatial reasoning. The algorithm for spatial inferences works in the following way. The initial interpretation of the first premise

The triangle is on the right of the circle

yields a propositional representation, which is constructed by a "compositional semantics":

((1 0 0) △ ○).

The parameters (1 0 0) specify which axes need to be incremented in order to relate the triangle to the circle (increment the right-left axis, i.e., keep adding 1 to it, as necessary; hold the front-back axis constant, i.e., increment it by 0; and hold the up-down axis constant, i.e., increment it by 0). There are no existing models of the discourse, because the assertion is first, and so a procedure is called that uses this propositional representation to build a minimal spatial representation:

○ △.

In the program, the spatial model is represented by an array. Likewise, the interpretation of the second premise

The circle is on the right of a line

yields the propositional representation

((1 0 0) ○ |).

This representation contains an item in the initial model, and so a procedure is called that uses the propositional representation to update this model by adding the line in the appropriate position:

| ○ △.

Given the further, third assertion

The triangle is on the right of the line, both items in its propositional representation occur in an existing model, and thus a procedure is called to verify the propositional representation. This procedure returns the value true, and with the proviso that the algorithm always constructs all possible models of the premises, the conclusion is therefore valid.

The algorithm has no need for a postulate capturing the transitivity of relations, such as "on the right of," which are emergent properties of the meaning of the relation and of how it is used to construct models. This emergence of logical properties has the advantage of accounting for a puzzling phenomenon—the vagaries in everyday spatial inferences. The inferences modeled in the program are for the "deictic" interpretation of "on the right of," that is, the relation as perceived from a speaker's point of view. Other entities have an intrinsic right-hand side and left-hand side, for example, human beings (see Miller and Johnson-Laird 1976, section 6.1.3). Hence the following premises:

Matthew is on Mark's right
Mark is on Luke's right

can refer to the position of three individuals in relation to the intrinsic right-hand sides of Mark and Luke. To build a model of the spatial relation, the inferential system needs to locate Mark, then to establish a frame of reference around him based on his orientation, and then to use the semantics of "on *X*'s right" to add Matthew to the model in a position on the right-hand side of the lateral plane passing through Mark (see Johnson-Laird 1983, 261). The same semantics as the program uses for "on the right" can be used, but instead of applying to the axes of the spatial array, it applies to axes centered on each individual according to their orientation. Hence, if the individuals are seated in a line, as in Leonardo da Vinci's painting of the Last Supper, then the model supports the transitive conclusion

Matthew is on Luke's right.

On the other hand, if they are seated round a small circular table, each premise can be true, but the transitive conclusion false. Depending on the size of the table and the number of individuals seated around it, transitivity can occur over limited regions, and the same semantics for "on *X*'s right" accounts for all the vagaries in the inference.

11.4 Experiment in Spatial Reasoning

The key feature of spatial models is not that they represent spatial relations—propositional representations also do that—but rather that they are functionally organized on spatial axes and, in particular, that information in them can be accessed by way of these axes. Does such an organization imply that when you have a spatial model of a situation, the relevant information will be laid out in your brain in a spatially isomorphic way? Not necessarily. A programming language, such as LISP, allows a program to manipulate spatial arrays by way of the coordinate values of their axes, but the data structure is only functionally an array and no corresponding physical array of data is necessarily to be found in a computer's memory as it runs the program. The same functional principle may well apply to high-level spatial models in human cognition.

The model theory makes systematically different predictions from those of theories based on formal rules. In an experiment reported by Byrne and Johnson-Laird (1989), the subjects carried out three sorts of spatial inference. The first sort were problems that could be answered by constructing just a single model of the premises, such as the following:

The knife is on the right of the plate.
The spoon is on the left of the plate.
The fork is in front of the spoon.
The cup is in front of the knife.
What's the relation between the fork and cup?

We knew from previous results that individuals tend to imagine symmetric arrangements of objects, and so these premises call for a model of this sort:

s	p	k
f		c

where *s* denotes a representation of the spoon, *p* a representation of the plate, and so on. This model yields the conclusion

The fork (*f*) is on the left of the cup (*c*).

There is no model of the premises that refutes this conclusion, and thus it follows validly from this single model of the premises. In contrast, if individuals reach this conclusion on the basis of a formal derivation, they must first derive the relation between the spoon and the knife. They need, for example, to infer from the second premise

The spoon is on the left of the plate

that the converse proposition follows:

The plate is on the right of the spoon.

They can then use the transitivity of "on the right of" to infer from this intermediate conclusion and the first premise that it follows that

The knife is on the right of the spoon.

At this point, they can use certain postulates about two-dimensional relations to derive the relation between the fork and the cup (see Hagert 1984 and Ohlsson 1984 for such formal rule systems of spatial inference).

Problems of the second sort yield multiple models because of a spatial indeterminacy, but they nevertheless support a valid answer. They were constructed by changing one word in the second premise:

The knife is on the right of the plate.
The spoon is on the left of the knife.
The fork is in front of the spoon.
The cup is in front of the knife.
What's the relation between the fork and cup?

The description yields models corresponding to two distinct layouts:

```
s   p   k
f       c

p   s   k
    f   c
```

Both these models, however, support the same conclusion:

The fork is on the left of the cup.

The model theory predicts that this problem should be harder than the previous one, because reasoners have to construct more than one model. In contrast, theories based on formal rules and propositional representations predict that this problem should be easier than the previous one because there is no need to infer the relation between the spoon and the knife: it is asserted by the second premise.

Problems of the third sort were similar but did not yield any valid relation between the two items in the question, for example:

The knife is on the right of the plate.
The spoon is on the left of the knife.
The fork is in front of the spoon.
The cup is in front of the plate.
What's the relation between the fork and cup?

In one of the experiments, eighteen subjects acted as their own controls and carried out the task with six problems of each of the three sorts presented in a random order. They drew reliably more correct conclusions to the one-model problems (70%) than to the multiple-model problems with valid answers (46%). Their correct conclusions were also reliably faster to the one-model problems (a mean of 3.1 seconds) than to the multiple-model problems with valid answers (3.6 seconds). It might be argued that the multiple-model problems are harder because they contain an irrelevant premise that plays no part in the inference. However, in an another experiment, the one-model problems contained an irrelevant premise, for example:

The knife is on the right of the plate.
The spoon is on the left of the plate.
The fork is in front of the spoon.
The cup is in front of the plate.
What's the relation between the fork and cup?

This description yields the following sort of model:

```
s   p   k
f   c
```

and, of course, the first premise is irrelevant to the deduction. Such problems were reliably easier (61% correct) than the multiple-model problems with valid conclusions (50% correct). Thus the results of the two experiments corroborate the model theory but run counter to theories that assume that reasoning depends on formal rules of inference.

11.5 Space for Time: Models of Temporal Relations

It seems entirely natural that human reasoners would represent spatial relations by imagining a spatial arrangement, but let us push the argument one step further. Perhaps spatial models underlie reasoning in other domains, that is, inferences that hinge on nonspatial matters may be made by manipulating models that are functionally organized in the same way as those representing spatial relations (see section 11.3). A plausible extrapolation is to *temporal* reasoning. Before we examine this extension, let us see how formal rules of inference might cope.

Formal rules might be used for temporal reasoning, but there are some obstacles to them. An obvious difficulty is the large variety of linguistic expressions, at least in Indo-European languages, that convey temporal information. Consider just a handful of illustrative cases. Verbs differ strikingly in their temporal semantics (see, for example, Dowty 1979; Kenny 1963; and Ryle 1949). For instance, the assertion "He was looking out of the window" means that for some interval of time at a reference time prior to the utterance the observer's gaze was out of the window. In contrast, the assertion "He was glancing out of the window" means that for a similar interval the observers gaze was alternately out of the window and not out of the window. Temporal adverbials can move the time of an event from the time of the utterance ("He is running now") to a time in the future ("He is running tomorrow"; see, for example, Bull 1963; Lyons 1977; and Partee 1984). General knowledge can lead to a sequential construal of sentential connectives, as in "He crashed the car and climbed out," or to a concurrent interpretation, as in "He crashed the car and damaged the fender." A theory of temporal language has to specify the semantics of these expressions, and particularly their contribution to the truth conditions of assertions. Formal rule theories of inference, in addition, must specify a set of inferential rules for temporal expressions.

In fact, no psychological theory based on formal rules of inference has so far been proposed for temporal reasoning, but logicians have proposed various analyses of temporal expressions. Quine (1974, 82) discusses the following pair of assertions:

I knew him before he lost his fortune
I knew him while he was with Sunnyrinse

and suggests treating them as assertions of the form, Some *F* are *G*, where *F* represents "moments in which I knew him" and *G* represents for the first assertion, "moments before he lost his fortune," and for the second assertion, "moments in which he was with Sunnyrinse." This treatment does not readily yield transitive inferences of the form

a before *b*,
b before *c*,
Therefore, *a* before *c*.

Other logicians have framed temporal logics as variants of modal logic (see, for example, Prior 1967; Rescher and Urquhart 1971), but these logics depend on simple temporal operators that do not correspond to the tense systems of natural language. Their scope is thus too narrow for the various forms of everyday expressions of time. Hence a more plausible way to incorporate temporal reasoning within a psychological theory based on formal rules of inference is to specify the logical properties of temporal expressions in "meaning postulates" in a way that is analogous to the psychological theories of spatial reasoning described in section 11.2.

Temporal relations probably cannot be imagined in a single visual image. In any case, the events themselves may not be visualizable, and manipulations of this factor have no detectable effects on reasoning (see, for example, Newstead, Manktelow, and Evans 1982; Richardson 1987; and Johnson-Laird, Byrne, and Tabossi 1989). When one imagines a temporal sequence, however, it often seems to unfold in time like the original events, though not necessarily at the same speed. This sort of representation uses time itself to represent the temporal axis (see Johnson-Laird 1983, 10). However, another possibility is to represent temporal relations in a static *spatial* model of the sequence of events in which one axis corresponds to time.

For example, the representation of the assertion

The clerk sounded the alarm after the suspect ran away

calls for a model of the form

```
r    a
```

in which the time axis runs from left to right, *r* denotes a representation of the suspect running away, and *a* denotes a representation of the clerk sounding the alarm. Events can be described as momentary or as having durations, definite or indefinite. Hence the further assertion

The manager was stabbed while the alarm was ringing

means that the stabbing occurred at some time between the onset and offset of the alarm:

```
r    a———
       s
```

where *s* denotes a representation of the stabbing, and the vertical dimension allows for contemporaneous events. This model corresponds to infinitely many different situations that have in common only the truth of the two premises. Thus the model contains no explicit representation of the duration for which the alarm sounded, or of the precise point at which the stabbing occurred. Yet, the conclusion

The stabbing occurred after the suspect ran away

is true in this model, and there is no model of the two premises that falsifies it.

I have implemented a computer program that carries out temporal inferences in exactly this way. It attempts to construct all the possible models of the premises. If the number grows too large, it then attempts to use the question—if there is one—to guide its construction of models so as to minimize the number it has to construct. Consider, for example, the following premises:

h happens before *b*
a happens before *b*
b happens before *c*
e happens before *d*
f happens before *d*
c happens before *d*
What's the relation between *a* and *d*?

When the program works through the premises in their stated order, it has to construct 239 models to answer the question—a number that vastly exceeds the capacity of human working memory. If the program's capacity is set more plausibly, say, to four models, it will give up working forwards and then try a depth-first search based on the question: What's the relation between *a* and *d*? It discovers the chain leading from the second premise (referring to *a*) through the third premise (referring to event *b*, which is also referred to by the second premise) to the final premise (referring to *d*), and constructs just the single model that these premises support. This model yields the conclusion that *a* happens before *d*. The advantages of this procedure are twofold. First, it ignores all irrelevant premises. Second, it deals with the premises in a coreferential order in which each premise after the first refers to an event already represented in the set of models. Of course, there are problems that defy the program's capacity for models even if it ignores irrelevant premises. In everyday life, however, individuals are unlikely to present information in an amount or in an order that overburdens human working memory; they are likely to be sensitive to the limitations of their audience (see Grice 1975). Hence it seemed appropriate in our experimental study of temporal reasoning to use similarly straightforward materials.

11.6 Experimental Study of Temporal Reasoning

Psychologists have not hitherto studied deductive reasoning based on temporal relations, and so Walter Schaeken, Gery d'Ydewalle (of the University of Leuven in Belgium), and the present author have carried out an series of experiments examining the topic.

Consider the premises of the following sort:

a before *b*
b before *c*
d while *a*
e while *c*
What's the relation between *d* and *e*?

where *a*, *b*, and so on stand for everyday events, such as "John shaves," "he drinks his coffee,"and so on. These events call for the construction of a single model:

a	b	c
d		e

where the vertical dimension allows for events to be contemporaneous. This model supports the conclusion

d before *e*.

The model theory predicts that this one-model problem should be easier than a similar inference that contains an indeterminacy. For example, the following premises call for several models:

a before *c*
b before *c*
d while *b*
e while *c*
What's the relation between *d* and *e*?

The premises are satisfied by the following models:

```
a   b   c        b   a   c        a   c
    d   e        d       e        b
                                  d   e
```

In all three models, *d* happens before *e*, and so it is a valid conclusion. The model theory also predicts that the time subjects spend reading the second premise, which creates the indeterminacy leading to multiple models, should be longer than the reading time of the second premise of the one-model problem. This multiple-model problem contains an irrelevant first premise, but the following one-model problem also contains an irrelevant first premise:

a before *b*
b before *c*
d while *b*
e while *c*
What's the relation between *d* and *e*?

In one of our experiments, we tested twenty-four university students with eight versions of each of the three sorts of problems above, and eight versions of a multiple-model problem that had no valid answer. The thirty-two problems were presented under computer control in a different random order to each subject. The two sorts of one model problem were easy and did not differ reliably (93% correct for the problems with no irrelevant premise and 89% correct for the problems with an irrelevant premise), but they were reliably easier than the multiple-model problems with valid conclusions (81% correct responses), which in turn were reliably easier than the multiple-model problems with no valid conclusions (44% correct responses). One would expect the latter problems to be difficult because it is vital to construct more than one model in order to appreciate that they have no valid conclusion, whereas the valid answer will emerge from any of the multiple models of the problems with a valid answer. Figure 11.1 shows the reading times for the four premises of the problems.

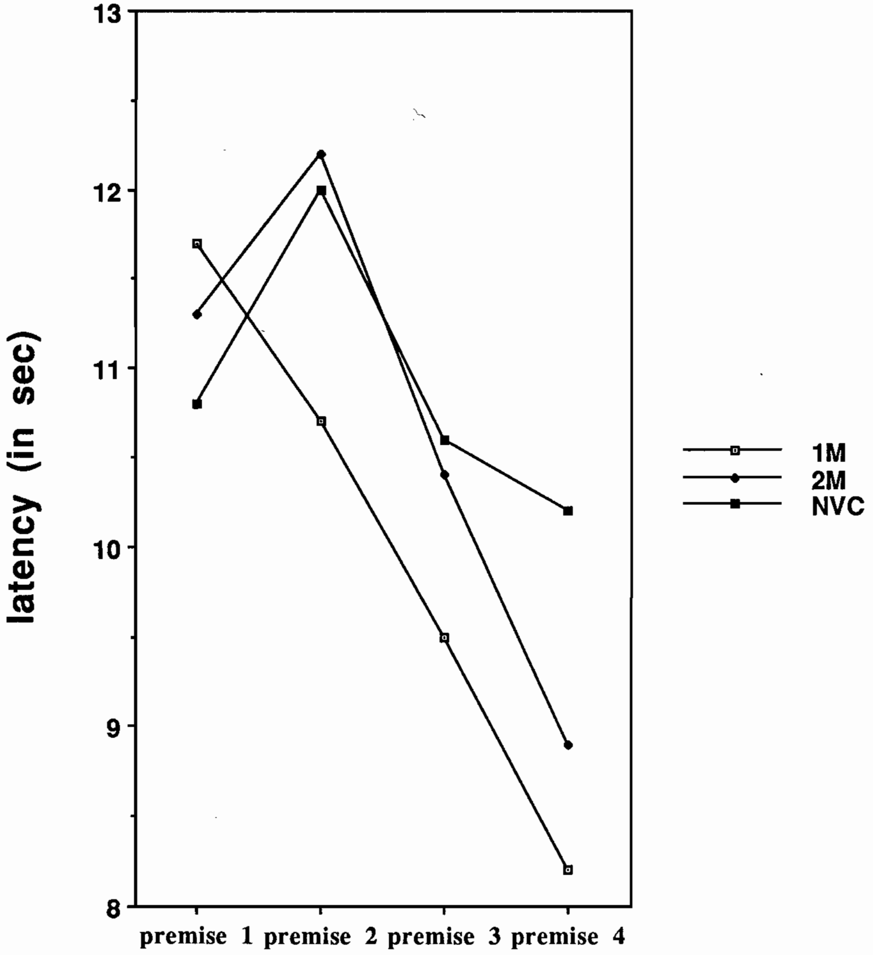


Figure 11.1

The mean latencies for reading the premises in the temporal inference experiment. The means are for one-model problems (1-M) collapsing over the two sorts, the multiple-model problems with a valid conclusion (2-M), and the multiple-model problems with no valid conclusion (NVC).

As the figure shows, subjects took reliably longer to read the second premise of the multiple-model problems—the premise that calls for the construction of more than one model—than to read the second premise of the one-model problems.

Our results, both for this experiment and others that we carried out, establish three main phenomena, and they imply that reasoning about temporal relations depends on mental models of the sequences of events. The first phenomenon concerns the number of models. When a description is consistent with just one model, the reasoning task is simple and subjects typically draw over 90% correct conclusions. When a description is consistent with more than one model, there is a reliable decline in performance. As in the earlier study of spatial reasoning, we pitted the predictions of the model theory against contrasting predictions based on formal rules of inference. The results showed that the one-model problems were reliably easier than the multiple-model problems, even though the one-model problems call for longer formal derivations than the multiple-model problems.

The second phenomenon concerns the subjects' erroneous conclusions. Formal rule theories make no specific predictions about the nature of such conclusions: subjects are said to err because they misapply a rule or fail to find a correct derivation. The model theory, however, predicts that erroneous conclusions arise because reasoners fail to consider all the models of the premises, and thus these conclusions should tend to be consistent with the premises (i.e., true in at least one model of them) rather than inconsistent with premises (i.e., not true in any model of them). The results corroborated this prediction of the model theory.

The third phenomenon concerns the time subjects took to read the premises and to respond to the questions. As we have seen, they took reliably longer to read a premise that led to multiple models than to read a corresponding premise in a one-model problem. Formal rule theories make no such prediction, and it is hard to reconcile this result with such theories because they make no use of models. The result also suggests that subjects do not construct models that represent indeterminacies within a single model. If they had done so, then they should have taken no longer to read these premises than the corresponding premises of one-model problems. And, of course, they should not have been more prone to err with indeterminate problems. The times to respond to the questions also bore out the greater difficulty of the multiple-model problems.

One final comment on our temporal experiments. Problems that depend on a transitive chain of events, as in the following one-model problem:

```
a   b   c
d       e
```

make an interesting contrast with one-model problems in which the transitive chain is not relevant to the answer:

```
a   b   c
    d   e
```

If subjects were imagining the events unfolding in time at a more or less constant rate, then presumably they ought to be able to respond slightly faster in the second case than in the first. That is to say, the actual temporal interval between *d* and *e* must be shorter in the second case than in the first. We examined this difference in the experiment described above. The mean latencies to respond were as follows: 7.0 seconds in the first case and 5.8 seconds in the second case. This difference was not too far from significance, and thus perhaps at least some of our subjects were imagining events as unfolding in time rather than simply constructing spatial models of the temporal relations.

11.7 Space for Space: How Diagrams Can Help Reasoning

Diagrams are often said to be helpful aids to thinking. They can make it easier to find relevant information—one can scan from one element to another element nearby much more rapidly than one might be able to find the equivalent information in a list of numbers or verbal assertions. Diagrams can make it easier to identify instances of a concept—an iconic representation can be recognized faster than a verbal description. Their symmetries can cut down on the number of cases that need to be examined. But can diagrams help the process of thought itself? Larkin and Simon (1987) grant that diagrams help reasoners to find information and to recognize it, but doubt whether they help the process of inference itself. According to Barwise and Etchemendy (1992, 82), who have developed a computer program, Hyperproof, that helps users to learn logic: "diagrams and pictures are extremely good at presenting a wealth of specific, conjunctive information. It is much harder to use them to present indefinite information, negative information, or disjunctive information. For these, sentences are often better." Hyperproof accordingly captures conjunctions in diagrams, but expresses disjunctions in verbal statements. The model theory, however, makes a different prediction. A major problem in deduction is to keep track of the possible models of premises. Hence a diagram that helps to make them explicit should also help people to reason. The result of perceiving such a diagram is a model—according to Marr's (1982) of vision—and thus one has a more direct route to a model than that provided by a verbal description. The verbal description needs to be parsed and a compositional semantics needs to be used to construct its propositional representation, which is then used in turn to construct a model. Hence it should be easier to reason from diagrams than from verbal descriptions.

We tested this prediction in two experiments based on so-called double disjunctions (Bauer and Johnson-Laird 1993). These are deductive problems, which are exemplified in verbal form by the following problem:

Julia is in Atlanta, or Raphael is in Tacoma, but not both.
Julia is in Seattle, or Paul is in Philadelphia, but not both.
What follows?

The model theory predicts that such problems based on exclusive disjunctions should be easier than those based on inclusive disjunctions:

Julia is in Atlanta, or Raphael is in Tacoma, or both.
Julia is in Seattle, or Paul is in Philadelphia, or both.
What follows?

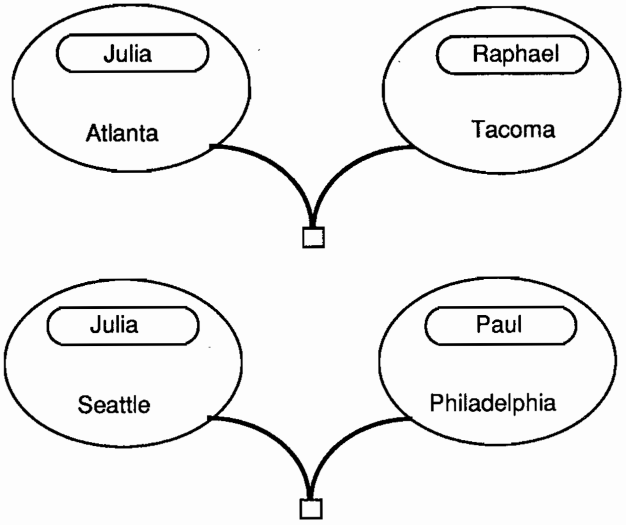
Each exclusive disjunction calls for only two models, whereas each inclusive disjunction calls for three models. Likewise, when the premises are combined, the exclusive problem yields three models:

a		p
s	t	
	t	p

Here *a* is a representation of Julia in Atlanta, *s* is a representation of Julia in Seattle, *t* is a representation of Raphael in Tacoma, and *p* is a representation of Paul in Philadelphia. In contrast, the inclusive problem yields a total of five models:

a		p
s	t	
	t	p
a	t	p
s	t	p

In our first experiment, premises of this sort were presented either verbally or else in the form of a diagram, such as figure 11.2. To represent, say, Julia in Atlanta, the diagram has a lozenge labeled "Julia" lying within the ellipse labeled "Atlanta." Inclusive disjunction, as the figure shows, is represented by a box connected by lines to the two component diagrams making up the premise as a whole. The experiment confirmed that exclusive disjunctions were easier than inclusive disjunctions (for both the percentages of correct responses and their latencies); it also confirmed that "identical" problems, in which the individual common to both premises was in the same place in both of them, were easier than "contrastive" problems such as the one above. But the experiment failed completely to detect any effect of diagrams: they yielded



What follows?

Figure 11.2
The diagrammatic presentation of double disjunctions in the first diagram experiment.

28% correct conclusions in comparison to the 30% correct for the verbal problems. Double disjunctions remained difficult, and these diagrams were no help at all.

With hindsight, the problem with the diagrams was that they used arbitrary symbols to represent disjunction and thus failed to make the alternative possibilities explicit. In a second experiment, we therefore used a new sort of diagram, as shown in figure 11.3, which is analogous to an electrical circuit. The idea, which we explained to the subjects, was to complete a path from one side of the diagram to the other by moving the shapes corresponding to people into the slots corresponding to cities. We tested four separate groups of subjects with logically equivalent problems: one group received diagrams of people and places (as in the figure); a second group received problems in the form of circuit diagrams of electrical switches; a third group received problems in the form of verbal premises about people and places; and a fourth group received problems in the form of verbal premises about electrical switches. There was no effect of the content of the problems—whether they were about people or switches—and therefore we have pooled the results. The percentages of correct responses are presented in figure 11.4. As the figure shows, there was a striking effect of mode of presentation: 74% correct responses to the diagrammatic problems in comparison to only 46% correct responses to the verbal problems. The

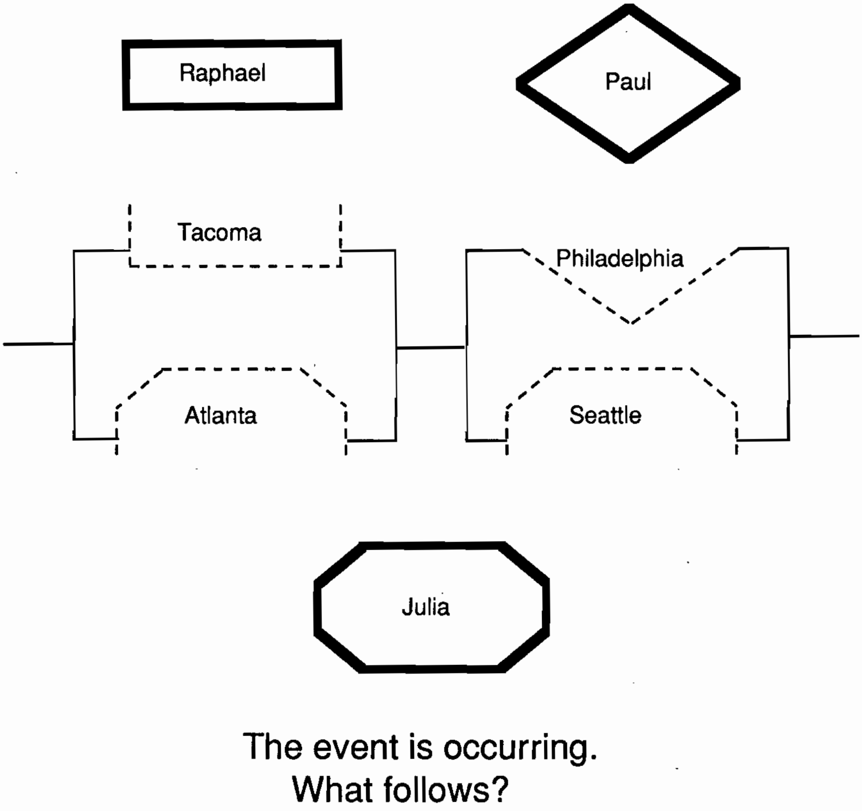


Figure 11.3
The diagrammatic presentation of double disjunctions in the second diagram experiment.

results also corroborated the model theory's predictions that exclusive disjunctions should be easier than inclusive disjunctions, and that identical problems should be easier than contrastive problems. The latencies of the subjects' correct responses had exactly the same pattern, for example, subjects were faster to reason with exclusive disjunctions than inclusive disjunctions, and they were reliably faster to respond to the diagrammatic problems (a mean of 99 seconds) than to the verbal problems (a mean of 135 seconds).

People evidently reason by trying to construct models of the alternative possibilities, and diagrams that enable these alternatives to be made explicit can be very helpful. With a diagram of the sort we used in our second experiment, individuals perceive the layout and in their mind's eye can move people into places and out again. By manipulating the model underlying the visual image, they can construct the alter-

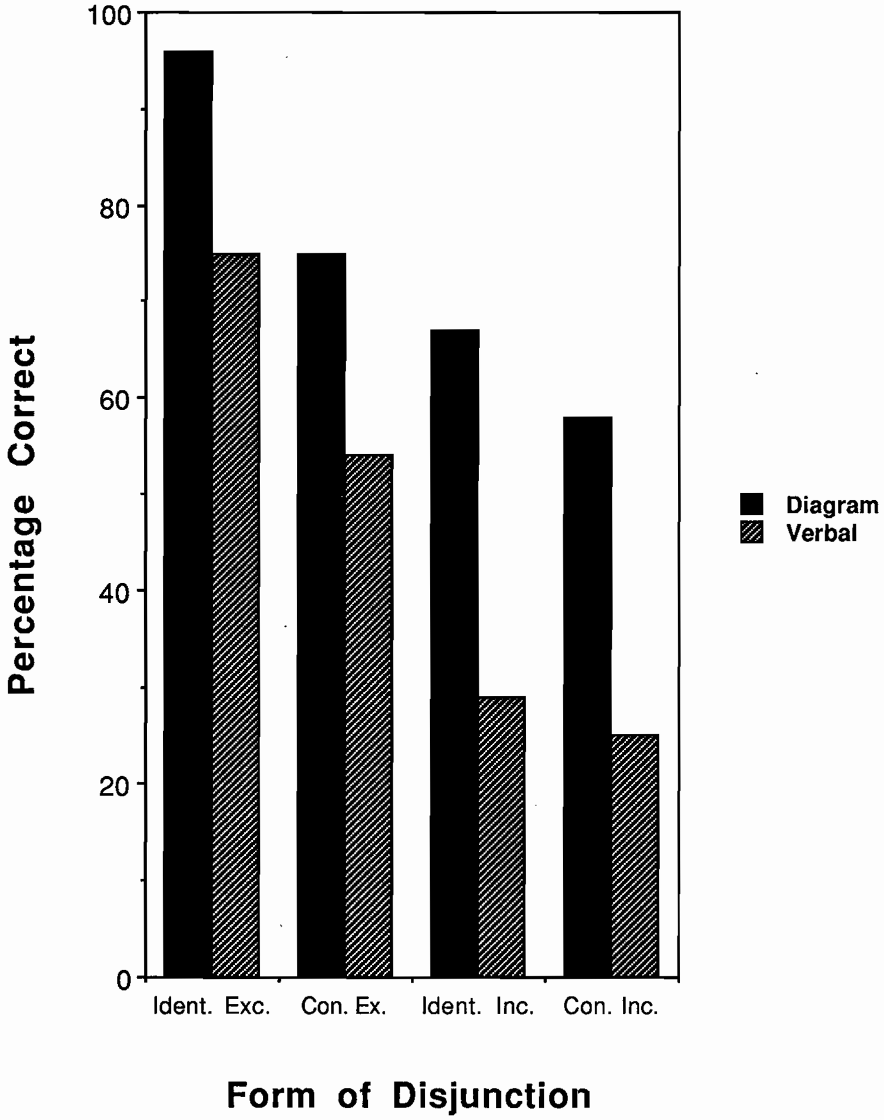


Figure 11.4
The percentages of correct responses in the second diagram experiment. There are two sorts of disjunction: exclusive (exc.) and inclusive (inc.), and two sorts of relation between premises: identical (ident.) and contrastive (con.).

native possibilities more readily than they can from verbal premises. It follows that diagrams are not merely encoded in propositional representations equivalent to those constructed from verbal premises (but see Baylor 1971, Pylyshyn 1973, and Palmer 1975 for opposing views).

11.8 Conclusions

Mental models are in many ways a primitive form of representation, which may owe their origin to the selective advantage of constructing internal representations of spatial representations in the external world. The evidence reviewed in this chapter suggests that mental models underpin the spatial reasoning of logically untutored individuals and may also play a similar role in temporal reasoning. Indeed, it may be that human inference in general is founded on the ability to construct spatial, or quasi-spatial models, which also appear to play a significant part in syllogistic reasoning and reasoning with multiple quantifiers (Johnson-Laird and Byrne 1991).

Historians of science and scientists themselves have often drawn attention to the role of diagrams in scientific thinking. Our studies show that not just any diagram has a helpful role to play. It is crucial that diagrams make the alternative possibilities explicit. Theories based on formal rules and propositional representations have to postulate the extraction of logical form from an internal description of visual percepts. In contrast, the model theory allows for inferences based on visual perception, which has a mental model as its end product (Marr 1982). The two theories accordingly diverge on the matter of diagrams. Formal rule theories argue that performance with a diagram should be worse than with the logically equivalent verbal premises: with a diagram, reasoners have to construct an internal description from which they can extract a logical form. The model theory, however, predicts that performance with a diagram that makes the alternative possibilities explicit should be better than with logically equivalent verbal premises: with a diagram, reasoners do not need to engage in the process of parsing and compositional semantics. The evidence indeed suggests that human reasoners use functionally spatial models to think about space, but they also appear to use such models in order to think in general.

Acknowledgments

I am grateful to Ruth Byrne for her collaboration in developing the theory of deduction based on mental models. I am also grateful to her, to Malcolm Bauer, and to Walter Schaeken for ideas and help in carrying out the present experiments. The research was supported in part by the James S. McDonnell Foundation.

References

Barwise, J. (1993). Everyday reasoning and logical inference. *Behavioral and Brain Sciences, 16*, 337–338. Commentary on Johnson-Laird and Byrne 1991.

Barwise, J., and Etchemendy, J. (1992). Hyperproof: Logical reasoning with diagrams. In N. H. Narayanan (Ed.), *AAAI Spring Symposium on Reasoning with Diagrammatic Representations*, 80–84. 25–27 March, Stanford University, Stanford, CA.

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

Baylor, G. W. (1971). Programs and protocol analysis on a mental imagery task. *First International Joint Conference on Artificial Intelligence*. N. P.

Bull, W. E. (1963). *Time, tense, and the verb.* Berkeley: University of California Press.

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

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

Dowty, D. R. (1979). *Word meaning and Montague grammar.* Dordrecht: Reidel.

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

Grice, H. P. (1975). Logic and conversation. In P. Cole and J. L. Morgan (Eds.), *Syntax and semantics.* Vol. 3: *Speech acts.* New York: Seminar Press.

Hagert, G. (1984). Modeling mental models: Experiments in cognitive modeling of spatial reasoning. In. T. O'Shea (Ed.), *Advances in artificial intelligence*, Amsterdam: North-Holland.

Johnson-Laird, P. N. (1975). Models of deduction. In R. Falmagne (Ed.), *Reasoning: Representation and process.* Hillsdale, NJ: Erlbaum.

Johnson-Laird, P. N. (1983). *Mental models: Toward a cognitive science of language, inference, and consciousness.* Cambridge, MA: Harvard University Press; Cambridge: Cambridge University Press.

Johnson-Laird, P. N. (1994). Mental models and probabilistic thinking. *Cognition*, 189–209.

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

Johnson-Laird, P. N., Byrne, R. M. J., and Tabossi, P. (1989). Reasoning by model: The case of multiple quantification. *Psychological Review, 96*, 658–673.

Kenny, A. (1963). *Action, emotion, and will.* New York: Humanities Press.

Kuhn, D. (1991). *The skills of argument.* Cambridge: Cambridge University Press.

Larkin, J., and Simon, H. (1987). Why a diagram is (sometimes) worth 10,000 words. *Cognitive Science, 11*, 65–99.

Lyons, J. (1977). *Semantics.* Vols. 1 and 2. Cambridge: Cambridge University Press.

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

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

Newstead, S. E., Manktelow, K. I., and Evans, J. St. B. T. (1982). The role of imagery in the representation of linear orderings. *Current Psychological Research, 2*, 21–32.

Ohlsson, S. (1984). Induced strategy shifts in spatial reasoning. *Acta Psychologica, 57*, 46–67.

Osherson, D. N., Smith, E. E., and Shafir, E. B. (1986). Some origins of belief. *Cognition, 24*, 197–224.

Palmer, S. E. (1975). Visual perception and world knowledge: Notes on a model of sensory-cognitive interaction. In D. A. Norman, D. E. Rumelhart, and the LNR Research Group (Eds.), *Explorations in cognition*, 279–307. San Francisco: Freeman.

Partee, B. (1984). Nominal and temporal anaphora. *Linguistics and Philosophy, 7*, 243–286.

Perkins, D. N., Allen, R., and Hafner, J. (1983). Difficulties in everyday reasoning. In W. Maxwell (Ed.), *Thinking.* Philadelphia: Franklin Institute Press.

Prior, A. N. (1967). *Past, Present, and Future.* Oxford: Clarendon Press.

Pylyshyn, Z. (1973). What the mind's eye tells the mind's brain: A critique of mental imagery. *Psychological Bulletin, 80*, 1–24.

Quine, W. V. O. (1974). *Methods of logic.* 3d ed. London: Routledge and Kegan Paul.

Rescher, N., and Urquhart, A. (1971). *Temporal logic.* New York: Springer.

Richardson, J. T. E. (1987). The role of mental imagery in models of transitive inference. *British Journal of Psychology, 78*, 189–203.

Ryle, G. (1949). *The concept of mind.* London: Hutchinson.

Schaeken, W., Johnson-Laird, P. N., and d'Ydewalle, G. (1994). Mental models and temporal reasoning. *Cognition*, in press.

Simon, H. A. (1959). Theories of decision making in economics and behavioral science. *American Economic Review, 49*, 253–283.

Toulmin, S. E. (1958). *The uses of argument.* Cambridge: Cambridge University Press.

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

Language and Space

edited by Paul Bloom, Mary A. Peterson, Lynn Nadel, and Merrill F. Garrett

A Bradford Book
The MIT Press
Cambridge, Massachusetts
London, England

This book was set in Times Roman by Asco Trade Typesetting Ltd., Hong Kong and was printed and bound in the United States of America.

Library of Congress Cataloging-in-Publication Data

Language and space / edited by Paul Bloom . . . [et al.].
p. cm. — (Language, speech, and communication)
Papers presented at a conference of the same name which was held
Mar. 16–19, 1994, Tucson, Ariz.
"A Bradford book."
Includes bibliographical references and index.
ISBN 0-262-02403-9
1. Space and time in language—Congresses. I. Bloom, Paul, 1963- . II. Series.
p37.5.S65L36 1996
401′.9—dc20 95-36427
CIP