How to reason syllogistically

In this chapter, I am going to describe a theory of syllogistic inference that my colleagues and I have gradually developed over the last seven years (see Johnson-Laird, 1975a; Johnson-Laird and Steedman, 1978; Johnson-Laird and Bara, 1982). It may seem surprising that it has taken so long to explain how people derive conclusions from what are, formally speaking, only sixty-four different pairs of premises. But the theory is, I believe, essentially correct; it is both descriptively and explanatorily adequate according to the criteria introduced in the previous chapter. Its various versions have been modelled in computer programs, and it has been corroborated by independent experimental results. It is in fact a special case of a more general theory of reasoning, but it is worth concentrating first on syllogisms because they straddle the boundary of human logical competence – some are literally child’s play while others are beyond the ability of all but the most expert reasoners.

The externalization of syllogistic inference

Let us begin by considering an imaginary method of externalizing the process of deduction. Suppose you want to draw a conclusion from the premises:

All the artists are beekeepers
All the beekeepers are chemists

without relying on Euler circles or Venn diagrams. One way in which to proceed is to employ a group of actors to construct a ‘tableau’ in which some of them act as artists, some as beekeepers, and some as chemists. To represent the first premise, every person acting as an artist is also instructed to play the part of a beekeeper, and, since the first premise is consistent with there being beekeepers who are not artists, that role is assigned to other actors, who are told that it is uncertain whether or not they exist. In short, a tableau of the following sort is set up:

\[
\begin{array}{c}
\text{artist} = \text{beekeeper} \\
\text{artist} = \text{beekeeper} \\
\text{artist} = \text{beekeeper} \\
(\text{beekeeper}) \\
(\text{beekeeper}) \\
\end{array}
\]

There are three actors playing the joint roles, and two actors taking the part of the beekeepers who are not artists – the parentheses designate a directorial device establishing that the latter may or may not exist. Obviously, the numbers of actors playing the different roles is entirely arbitrary. The tableau is easily extended to accommodate the second premise:

All the beekeepers are chemists.

Those actors playing beekeepers are instructed to take on the role of chemists, and an arbitrary number of new actors are introduced to play the role of chemists who are not beekeepers – a type which, once again, may or may not exist:

\[
\begin{array}{c}
\text{artist} = \text{beekeeper} = \text{chemist} \\
\text{artist} = \text{beekeeper} = \text{chemist} \\
\text{artist} = \text{beekeeper} = \text{chemist} \\
(\text{beekeeper}) = (\text{chemist}) \\
(\text{beekeeper}) = (\text{chemist}) \\
\end{array}
\]

At this point, if you were asked whether it followed that All the artists are chemists, you could readily inspect the tableau and determine that this conclusion is indeed true. All that you have done is to externalize and to combine the information in two premises. Deductive inference, however, requires more than just the construction of an integrated representation of the premises; it calls for a search for counter-examples. A more complicated example will illustrate this point.

Consider how a troupe of actors could represent the premises:

None of the authors are burglars.
Some of the chefs are burglars.

The first premise is represented by two distinct groups acting as the authors and the burglars, and they are instructed that they are never allowed to
MENTAL MODELS

take on each other’s role. The two groups are, as it were, fenced off from
each other:

author

author

author

burglar

burglar

burglar

The tableau indicates that no author is identical to any burglar. The
information in the second premise, Some of the chefs are burglars, is added
by extending the tableau in a straightforward way, where some is taken to
mean at least some:

author

author

author

burglar = chef

burglar = chef

(burglar) (chef)

It is tempting, at this point, to conclude that None of the authors are chefs,
or conversely, that None of the chefs are authors. (Six subjects out of twenty
drew conclusions of this form in an experiment reported by Johnson-Laird
and Steedman, 1978.) Both conclusions are certainly consistent with the
tableau. But neither is warranted, because there is another way of represent-
ing the premises:

author

author

author = chef

burglar = chef

burglar = chef

(burglar) (chef)

You can check for yourself that this interpretation is wholly consistent
with the premises (None of the authors are burglars, and Some of the chefs
are burglars), yet it invalidates the previous conclusions and suggests
instead that Some of the authors are not chefs. However, there is still another

HOW TO REASON SYLLOGISTICALLY

possibility. There could be several chefs who are not burglars, and each
author could be identical to such a chef:

author = chef

author = chef

author = chef

burglar = chef

burglar = chef

(burglar)

The tableau is still consistent with the premises, but what it shows is that
the conclusion that Some of the authors are not chefs is invalid. There is
no other assignment of roles that is compatible with the premises, and
you might therefore suppose that there is no valid conclusion (as did six
subjects in the experiment). But in all three of the tableaux it is the case
that at least Some of the chefs are not authors. This conclusion is accordingly
valid. (It was drawn by seven subjects.)

An effective procedure for syllogistic inference

The idea of employing actors to take on different parts suggests, of course,
a hypothesis about how people might actually make inferences. Instead of
arranging an external tableau, they could construct a mental model – an
internal tableau containing elements that stand for the members of sets in
just the same way that the actors did. A general procedure for making
inferences in this way requires three main steps.

1. Construct a mental model of the first premise. The representation of a
universal affirmative assertion has the following structure:

All of the X are Y: x = y

x = y

(y)

(y)

where the number of tokens corresponding to x’s and y’s is arbitrary, and
the items in parentheses represent the possible existence of y’s that are
not x’s. The representations of the other forms of syllogistic premise are
MENTAL MODELS

Some of the X are Y:
\[ x = y \]
\[ (x) (y) \]

None of the X are Y:
\[ x \]
\[ \underbrace{\ldots\ldots\ldots\ldots\ldots\ldots} \]
\[ y \]

Some of the X are not Y:
\[ x \]
\[ \underbrace{\ldots\ldots\ldots\ldots\ldots\ldots} \]
\[ (x) = y \]
\[ y \]

It should be noted that each premise requires only a single mental model.

A crucial point about mental models is that the system for constructing and interpreting them must embody the knowledge that the number of entities depicted is irrelevant to any syllogistic inference that is drawn. In the case of numerical or proportional inferences, however, the numbers or proportions will matter. The procedure must accordingly have an independent way of keeping track of which particular proposition is being modelled. Human reasoners appear to retain a superficial representation of the propositions expressed by the premises – one that is close to their linguistic form – but from the errors they make, they appear to make inferences by manipulating mental models rather than by deploying rules of inference on these superficial representations.

2. Add the information in the second premise to the mental model of the first premise, taking into account the different ways in which this can be done. In addition to the purely interpretative skills required to construct mental models, reasoners must appreciate the fundamental semantic principle underlying valid deduction: an inference is valid if and only if there is no way of interpreting the premises that is consistent with a denial of the conclusion. This principle motivates the search for alternative ways of adding the information from the second premise.

For some syllogisms, there is only one possible integrated model. For example, a premise of the form Some of the A are B yields the model:
\[ a = b \]
\[ a = b \]
\[ (a) (b) \]

and a second premise of the form All of the B are C can be integrated only by forming the model:
\[ a = b - c \]
\[ a = b = c \]
\[ (a) (b) = c \]
\[ (c) \]

from which it follows that Some of the A are C. There is no alternative model of the premises that violates this conclusion.

For other syllogisms, it is necessary to construct and evaluate two models. For example, premises of the form:

Some of the A are not B
All the C are B

can be represented by the model:
\[ \]
\[ a \]
\[ \underbrace{\ldots\ldots\ldots\ldots\ldots\ldots} \]
\[ (a) = b = c \]
\[ b = c \]
\[ (b) \]

In this model, the conclusion Some of the A are not C is true because the A’s above the broken line are not C’s, and the converse conclusion Some of the C are not A is also true because there is at least one C that is not linked to an A. The search for an alternative model to falsify these putative conclusions yields a second model:
\[ a \]
\[ \underbrace{\ldots\ldots\ldots\ldots\ldots\ldots} \]
\[ (a) = b = c \]
\[ (a) = b = c \]
\[ (b) \]

This model rules out the second conclusion because all the C’s are A’s, but there is no way to destroy the first conclusion, Some of the A are not C, which is accordingly valid.

Still other premises yield three different models. For example, the premises:

All of the B are A
None of the B are C

99
MENTAL MODELS

yield the model:

\[
\begin{array}{c}
c \\
\hline
b = a \\
b = a \\
(a)
\end{array}
\]

which suggests the conclusion *None of the C are A*, or its converse *None of the A are C*. The search for a counter-example yields the model:

\[
\begin{array}{c}
c \\
\hline
c = a \\
b = a \\
b = a \\
(a)
\end{array}
\]

which falsifies both of these conclusions and suggests instead that *Some of the C are not A* or conversely that *Some of the A are not C*. The search for a counter-example to these conclusions yields:

\[
\begin{array}{c}
c = a \\
\hline
c = a \\
b = a \\
(a)
\end{array}
\]

which shows that only the conclusion *Some of the A are not C* is valid. The last model by itself, of course, suggests the invalid conclusion that *All of the C are A*, but that conclusion is ruled out by the previous models. Although these premises have three models, they do not all have to be constructed *ab initio*: an entirely feasible strategy, as illustrated here, is to construct one model and then to try out various modifications of it that are consistent with the premises.

Since the procedure is based on the assumption that each of the four forms of syllogistic premise is represented by just one mental model, it avoids the explosion of combinatorial possibilities that so embarrasses the theories based on Euler circles. The assumption is feasible provided that optional entities are directly represented within a model, as in the representation of *All A are B* as:

\[
\begin{array}{c}
a = b \\
a = b \\
(b)
\end{array}
\]

HOW TO REASON SYLLOGISTICALLY

Is it logically possible to extend this notion so that instead of the three models required by the last example just one model suffices? Such a model might have the following form:

\[
\begin{array}{c}
c = (a) \\
\hline
c = (a) \\
b = a \\
b = a \\
(a)
\end{array}
\]

However, its correct interpretation requires a new notational principle to be introduced in order to make plain that both, one, or neither of the identities above the broken line may apply. Such a device is merely a notational variation on the present theory: it is still necessary to consider three logically distinct possibilities corresponding to all, some, or none of the c’s being identical to a’s.

Syllogisms never require more than three different mental models to be constructed, and Table 5.1 presents the number of models that are required for each of the twenty seven pairs of premises that yield a non-trivial valid conclusion. The table indicates when such a conclusion is contrary to the figural effect, and should therefore be harder to draw. It also presents the percentages of correct responses made by subjects in an experiment designed to test the theory (see Johnson-Laird and Bara, 1982). Every subject was presented with each of the sixty-four possible premises with a different sensible content and asked to state what conclusion, if any, followed from the premises. The results will be analysed later in the chapter together with those from other experiments.

3. *Frame a conclusion to express the relation, if any, between the 'end' terms that holds in all the models of the premises.* An 'end' term is one which occurs in only a single premise, unlike the 'middle' term which occurs in both premises. If there is no such relation between the end terms, the only valid conclusions that can be drawn are trivial ones, such as a conjunction or disjunction of the premises, and subjects generally respond that there is no valid conclusion.

The difficulty of syllogistic inference depends on the number of mental models of the premises

If every step of the procedure is carried out correctly, then as a computer implementation of it (in POP-10) shows, the result is a completely rational
and error-free performance. A mistake in any step may, or may not, vitiate the conclusion. Since the first and third steps are common to the comprehension and production of much discourse, they are likely to be relatively free of error. The task of drawing a valid conclusion should be relatively easy when there is only one possible integrated model; even an individual who does not appreciate the need to search for counter-examples should make the correct response. The major source of difficulty should be the construction and evaluation of alternative models. This process places an additional load on working memory – the memory system that is used for holding the mental models while they are manipulated. The greater the number of models to be considered, the harder it should be to construct and to evaluate them. Hence, the task should be harder when it is necessary to construct two models, and harder still when it is necessary to construct three models.

Working with various colleagues – principally Janelen Huttenlocher, Kate Ehrlich, and Bruno Bara – I have carried out a number of experiments on inference in which subjects have been asked to state what, if anything, follows from a series of syllogistic premises. We have used premises with sensible contents, but which do not predispose the subjects to any particular conclusions on a factual basis. Table 5.2 summarizes the results of three of our experiments. In experiment 1, twenty students at Teachers College, Columbia University were given all sixty-four possible pairs of premises. In experiment 2, twenty students at Milan University were also given the sixty-four pairs of premises. And in experiment 3, a further twenty Milanese students were given the same problems but had only ten seconds to respond to each of them. The table shows the percentages of correctly drawn valid conclusions as a function of the number of mental models that have to be constructed to make the right response for the right reason. Despite the variety of conditions and subjects in the experiments, there was a highly reliable trend in each of them: the greater the number of models to be considered, the harder the task of drawing a valid conclusion.

Subjects do indeed attempt to assess alternative models of the premises, but often the task exceeds the capacity of their working memories. It is difficult to hold in mind one model whilst attempting to construct and to evaluate an alternative model. This conclusion is corroborated by the nature of the errors that the subjects make. They almost invariably (i.e., 90% or more) respond with a conclusion that is congruent with only some of the possible models of the premises.

The figural effects

There is one other major factor that affects syllogistic inference: the figure of the premises. The figural effects consist in a bias towards certain forms of conclusion, and an increase in the difficulty and latency of drawing correct conclusions over the four figures. These phenomena are important because they cannot be readily explained in terms of atmosphere, illicit conversion, or theories based on Euler circles and Venn diagrams. But they can be accounted for in terms of the processes that occur in forming an integrated mental model of premises. If the two occurrences of the middle term are in a figure that readily permits the two premises to be integrated, the task is relatively easy; if they are in a figure that does not permit an immediate integration, then additional operations have to be carried out to make it possible, and the premises that require these additional operations are harder to integrate. These additional operations have been modelled in the computer program too. The fundamental assumption on which they are based is that working memory operates on a 'first in, first out' basis. For example, it is easier to recall a list of digits in the order in which they were presented – the first digit recalled first, the second digit recalled second, and so on – than in the opposite order (see Broadbent, 1958, p. 236). The same principle applies to syllogisms, and the natural order in which to state a conclusion is the order in which the terms were used to construct a mental model of the premises.
MENTAL MODELS

With premises in the $A - B$, $B - C$ figure, the two instances of the middle term, $B$, occur one after the other, and it is easy to construct a mental model of the first premise and then immediately to integrate the information from the second premise. For example, with premises of the form:

Some of the $A$ are $B$
All of the $B$ are $C$

a reasoner can form a model of the first premise:

$$a = b$$
$$a = b$$

(a) (b)

and then immediately integrate the content of the second premise by substituting $c$'s for $b$'s:

$$a = c$$
$$a = c$$

(a) (c)

This procedure of substituting one type of token for another was deliberately omitted in the account of the theory earlier in the chapter, but it is an essential part of the explanation of the figural effects, because the substitution requires an appropriate ordering of the terms within the premises: there has to be temporal contiguity between the original items ($b$'s) and the ones that replace them ($c$'s). Since the $a$'s preceded the $c$'s into working memory in the example, the 'first in, first out' principle leads to a conclusion of the form Some of the $A$ are $C$ rather than its equally valid converse. The same principle applies to any premises in the $A - B$, $B - C$ figure, and thus favours conclusions of the form $A - C$ rather than $C - A$ (cf. Table 5.3).

With premises in the $B - A$, $C - B$ figure, the two instances of the middle term do not occur one after the other. They are separated and therefore the process of substitution cannot occur. A natural way to proceed, however, is to construct a model of the second premise, $C - B$, renew the interpretation of the first premise, $B - A$, and then substitute the information it contains in the model of the second premise. The reordering of the premises brings together the two occurrences of the middle term. For example, with premises of the form:

All of the $B$ are $A$
Some of the $C$ are $B$

HOW TO REASON SYLLOGISTICALLY

reasoners are unable to make an immediate substitution, because of the separation of the two occurrences of the middle term. They accordingly construct a model of the second premise:

$$c = b$$
$$c = b$$

(c) (b)

make a renewed interpretation of the first premise, and substitute its information in the model:

$$c = a$$
$$c = a$$

(c) (a)

(a)

On the 'first in, first out' principle of working memory, the resulting mental model will yield a conclusion of the form Some of the $C$ are $A$. In general, premises in this figure should produce a bias towards $C - A$ conclusions (cf. Table 5.3).

The two remaining figures are still more complicated. There are two possible routes by which to integrate premises in the $A - B$, $C - B$ figure. Reasoners can construct a model of the first premise, $A - B$, and then switch round the order of the terms in their interpretation of the second premise to $B - C$ so that the two instances of the middle term occur one after the other. Alternatively, they can construct a model of the second premise, $C - B$, renew their interpretation of the first premise, then switch it round to $B - A$ so as to make the substitution possible. Switching round an interpretation must not be confused with the operation of converting a premise, though the two notions are similar. The converse of Some $A$ are $B$ is Some $B$ are $A$ and they are equivalent in that when one is true the other is true; the converse of All $A$ are $B$ is All $B$ are $A$, and they are not equivalent. If reasoners simply formed the converse of an expression, they would often fall into error. Although various theorists have argued that illicit conversions are a source of error (see Chapter 4), the idea of switching round an interpretation concerns only the order of information in working memory. The interpretation of All the $A$ are $B$ takes the form:

$$a = b$$
$$a = b$$

(b)

(b)
MENTAL MODELS

If this interpretation is switched round, it takes the form:

\[
\begin{align*}
 b &= a \\
 a &= b \\
(b) \\
(b)
\end{align*}
\]

This revision is logically accurate: an illicit conversion would only occur if the tokens representing the possibility of b's that were not a's were dropped in the process. The purpose of the operation is to ensure that the two instances of the middle term occur one after the other.

In the remaining figure, B – A, B – C, there are again two alternative procedures. Reasoners may switch round a model of the first premise to A – B, form a model of it, and then substitute the information from the second premise, B – C, in the model. Alternatively, they may switch round a model of the second premise to C – B, renew their interpretation of the first premise, B – A, and substitute its information in the model.

The operations that are required to form a mental model from premises in the four figures are summarized in Table 5.3. The bias they create towards one particular form of conclusion for the asymmetric figures, A – B and B – A, is highly reliable. The slight bias towards A – C conclusions for the symmetric figures, A – B and B – A, probably reflects a preference both for constructing an initial model from the first premise and for a process that requires fewer operations.

Two basic operations are required to construct a model of the premises in the A – B, figure: building an initial model and integrating the information from the second premise. A third operation of renewing the interpretation of a premise is required for premises in the B – A, figure. A more complex operation is required by both procedures for premises in the A – B, figure: switching round the interpretation of a premise so that information about its end term can be substituted in the model. Finally, a still more complex operation is required by both procedures for premises in the B – C, figure:

<table>
<thead>
<tr>
<th>Operations</th>
<th>Switch round model</th>
<th>Build model</th>
<th>Renew interpretation</th>
<th>Integrate</th>
<th>Predicted response bias</th>
<th>Obtained response bias</th>
</tr>
</thead>
<tbody>
<tr>
<td>Experiment 1</td>
<td>2.8%</td>
<td>80.1%</td>
<td>19.1%</td>
<td>90%</td>
<td>76%</td>
<td>74%</td>
</tr>
<tr>
<td>Experiment 2</td>
<td>2.8%</td>
<td>80.1%</td>
<td>19.1%</td>
<td>90%</td>
<td>76%</td>
<td>74%</td>
</tr>
<tr>
<td>Experiment 3</td>
<td>2.8%</td>
<td>80.1%</td>
<td>19.1%</td>
<td>90%</td>
<td>76%</td>
<td>74%</td>
</tr>
</tbody>
</table>
switching round the interpretation of a premise as a whole as a prerequisite for constructing an initial model.

Since the complexity of the operations required to form a mental model increases over the four figures, there should be a corresponding increase in the difficulty of drawing a valid conclusion, and in the latency of responses. An experiment carried out by Johnson-Laird and Bara (1982) was specifically designed to test these predictions (see Experiment 2 of Table 5.2). There was a reliable decline in the percentages of correct conclusions over the four figures: 51%, 48%, 35%, and 22% respectively. As Table 5.2 shows, the percentages of correct responses where three models had to be constructed were too small to allow us to obtain reliable estimates of latency, while premises requiring two models only occur in the two symmetric figures and hence their data did not enable us to test the latency prediction. However, the trend in latencies for the premises requiring one model reliably confirmed the prediction; over the four figures, the means were 11.6, 12.9, 18.7, and 22.1 seconds respectively.

An alternative explanation of the figural effects

In principle, there are many other possible explanations for the effects of figure on the form of conclusions, their accuracy, and their latency. I believe, however, that the present theory is correct in its essentials, and to substantiate this claim I shall consider an alternative account and some further experimental results.

Many psychologists confronted with the figural effects assume that they somehow reflect the order of the premises and their 'given - new' structure, that is, the position of the given information and the new information in the premises. In fact, this superficially appealing hypothesis fails to account for the phenomena. It is certainly true that the optimal ordering of given and new information corresponds to the figure:

\[
A \rightarrow B \\
B \rightarrow C
\]

with the given information in the second premise, B, referring back to the most recent new information in the first premise. The hypothesis predicts that the other figures, which violate this optimal ordering, should be harder to understand, but it does not explain the trend in difficulty over the remaining figures. Likewise, it does not account for the bias in the form of the conclusions.

A further difficulty for the 'given - new' explanation of the figural effects derives from a series of experiments reported by Ehrlich and Johnson-Laird (1982). In one condition of these experiments, the subjects had to remember premises of the form:

A is on the right of B  
B is in front of C  
C is on the left of D

in order to draw a corresponding diagram. These experiments, which will be reported in Chapter 14, were primarily concerned with quite a different matter, but one of the independent variables was the figural arrangement of the terms, A, B, C, D, which in fact designated common objects. Figure had no consistent effects on reading times or on memory for the premises, as reflected in the accuracy of the diagrams. It therefore seems that figural effects occur primarily when an inference has to be made, and in particular when a direct link has to be established between the end terms - with the middle term dropping out of the final representation of the conclusion.

If figural effects are a consequence of the operations required to make inferences, then they should occur in all forms of deduction. There should be figural effects, for example, in simple three-term series problems of the form:

Alice is taller than Bertha.  
Bertha is taller than Carol.

Indeed, in his classic paper, Ian Hunter (1957) proposed that the difficulty of such problems exactly reflected their figural arrangement. Yet, despite the many studies of three-term problems, there had been no investigation of the conclusions that subjects spontaneously state in their own words until Bruno Bara, Patrizia Tabossi and I carried out two experiments using this procedure in order to test whether there was a figural bias on such conclusions.

One factor that affects the difficulty of three-term series problems is that there is, as Herb Clark has established, a difference in the ease of understanding such pairs of expressions as taller than and shorter than (see Clark and Clark, 1977). Taller than is a neutral expression that implies nothing about the absolute heights of the entities it relates, whereas shorter than suggests that these entities are short rather than tall. This contrast between the neutral or ‘unmarked’ term and its ‘marked’ antonym is clearest in the difference between such questions as ‘Which of the two is the taller?’ and ‘Which of the two is the shorter?’, and in the fact that tallness rather
Mental Models

than shortness is the name of the dimension as a whole. In general, unmarked terms are easier to grasp than marked terms. Hence, in order to obviate this factor in our experiments, we used problems in which the relational term is its own converse:

Alice is related to Bertha
Bertha is related to Carol

where it was clear to the subjects that the premises were about blood-relationships. The experiments showed that there was a general bias towards A – C conclusions, and that, as predicted, premises in the A – B, B – C figure enhanced this bias (77% A – C conclusions), whereas those in the B – A, C – B figure eliminated it (only 47% A – C conclusions).

Evidently, the figural effect is not peculiar to syllogisms, but truly mirrors the task of combining premises in working memory. The present theory seems to give a comprehensive account of it.

Syllogisms that yield no valid conclusions interrelating end terms

There are thirty-seven out of the sixty-four pairs of syllogistic premises for which there are no valid conclusions interrelating end terms. In principle, their correct evaluation requires a reasoner to discover that the alternative models of the premises have nothing in common, but the right response may sometimes be made for the wrong reason.

If both premises are particular, i.e., contain the quantifier some, then they can be interpreted by alternative models that support contradictory conclusions. For example, the premises:

Some of the A are B
Some of the B are C

are readily interpretable both where identities are maximized:

\[ a = b = c \]
\[ a = b = c \]

(a) (b) (c)

and where they are not:

\[ a = b \]
\[ a = b \] (c)

(a) b = c

b = c

112

Hence, it is readily apparent that such premises do not support a valid conclusion interrelating the end terms. When both premises are negative, they likewise support blatantly inconsistent models. For example, the premises:

None of the A are B
None of the B are C

yield both the model:

\[ a \]
\[ b \]

\[ a = c \]

\[ b \]

\[ c \]

and the model:

\[ a = c \]

\[ a = c \]

\[ b \]

in which there is no relation between the end items common to both interpretations.

The remaining premise pairs require three models to be constructed if the correct response of 'No valid conclusion' is to be guaranteed. For example, the premises:

All of the A are B
Some of the B are not C

yield a model:

\[ a \]

\[ a \]

\[ (c) \]

\[ c \]

which suggests that None of the A are C, or its converse. They also yield another model:

\[ a \]

\[ a = c \]

\[ c \]

which rules out the previous conclusions and suggests instead that Some
MENTAL MODELS

of the A are not C or its converse. These conclusions are only falsified by constructing a third model of the premises:

\[ b \]
\[ a = c \]
\[ a = c \]

The second and third models taken together suggest that at least Some of the A are C, or its converse; these conclusions are eliminated only by bearing in mind the first model.

In summary, the theory establishes three main categories of problem for which there is no valid conclusion interrelating the end terms. There are no grounds for supposing that it should be easier to construct two inconsistent models from two particular premises than from two negative premises, or vice versa. The two sorts of problem require independent skills. What can be predicted, however, is that both should be easier than problems that definitely require three models to be constructed to refute putative conclusions. As Table 5.4 indicates, the results of the experiments reliably confirmed this prediction.

Table 5.4 The percentages of correct responses to premises with no valid conclusions in three experiments. The percentages are shown as a function of the number of models that have to be constructed in order to guarantee the correctness of the response.

<table>
<thead>
<tr>
<th>Premises requiring two models</th>
<th>Premises requiring three models</th>
</tr>
</thead>
<tbody>
<tr>
<td>Experiment 1</td>
<td>68</td>
</tr>
<tr>
<td>Experiment 2</td>
<td>46</td>
</tr>
<tr>
<td>Experiment 3</td>
<td>71</td>
</tr>
</tbody>
</table>

Inference and working memory

The theory of mental models implies that two principal factors affect the difficulty of making an inference: the number of models to be constructed, and the figural arrangement of terms within the premises. As the results of our experiments show, both factors affect performance highly significantly, but they do not interact. Their effects do add up, however, to produce a peculiarly difficult variety of syllogism. Four pairs of premises out of the sixty-four require three models to be constructed that have in common only a conclusion that runs counter to the 'first in, first out' principle underlying the figural effect. Here is an example of such a problem, which the reader may recall from the previous chapter:

All of the bankers are athletes.
None of the councillors are bankers.

There are three ways of integrating the premises in a mental model:

\[ a \]
\[ a = a \]
\[ a = a \]
\[ (a) \]
\[ c = a \]
\[ c = a \]
\[ c = a \]

The first model suggests the conclusion:

None of the councillors are athletes

which was drawn by twelve subjects in Experiment 1 – the experiment that yielded the best overall performance. Two subjects drew the converse conclusion. The second model shows that both these conclusions are invalid and suggests instead:

Some of the councillors are not athletes

which was drawn by a further two subjects. The third model shows that this conclusion is invalid, too. Four subjects may have succeeded in constructing this model, since they responded that there was no valid conclusion. Not a single subject, however, appreciated that there is a conclusion that covers all three models, though it runs counter to the figure of the premises:

Some of the athletes are not councillors.

These results are characteristic for the four most difficult syllogisms of all – those that run counter to figure and require three models to be evaluated.

The effects of both number of models and figure arise from an inevitable bottleneck in the inferential machinery: the processing capacity of working memory, which must hold one representation in a store, while at the same time the relevant information from the current premise is substituted in it. This problem is not obviated by allowing subjects to have the written premises in front of them throughout the task: the integration of premises has to occur in working memory, unless the subjects are allowed to use paper and pencil so as to externalize the process.

The effect of number of models on inferential performance, so amply confirmed by the studies of syllogisms, is also detectable in other sorts of
Mental Models

In an experiment carried out by Johnson-Laird and Wason (1970) the task was to check whether a description of the contents of an envelope was correct. The subjects selected diagrams one at a time from a set laid out in front of them, and they were told by the experimenter whether or not each such diagram was in the envelope. The subjects could use this information to determine whether the description was true or false. The logically prudent strategy in this task is to concentrate on diagrams that do not fit the description on the envelope: if such a diagram is in the envelope, plainly the description is false. Some subjects, however, choose diagrams that fit the description. That choice is uninformative once it is known that the envelope contains something, because there is no reason why a diagram that fits the description should not also be outside the envelope. What the experiment showed was that a complex disjunctive description:

There is a dot which is not connected to any dot or every dot is connected to every dot

had a striking effect on the subjects' insight into the task. A subject would perform perfectly with other descriptions, only to lose that insight on the very next trial when the disjunctive description occurred. The effect did not seem to be merely a function of the presence of a negation or of several quantifiers, because the subjects were able to cope with such descriptions as:

There is a dot connected to a dot to which no other dot is connected.

The crucial factor seemed to be that this description could be represented in a single mental model of a prototypical relation:

A • •

B • •

where the required dot, A, is connected to another dot, B, that has no other connections. The disjunctive description, however, contains two mutually exclusive states of affairs:

There is a dot which is not connected to any dot or:

Every dot is connected to every dot.

Hence, it can be represented only by keeping in mind a disjunction of two alternative prototypical relations: which satisfies the first disjunct, and:

which satisfies the second. As Wason and I wrote:

it is possible that this [disjunctive description] occupies a greater amount of short-term memory than a single complex rule, and thus leaves a smaller amount of 'computing space' available for handling the selection of the diagrams. (Johnson-Laird and Wason, 1970, p. 58)

More recently, Baddeley and his colleagues have made a comprehensive examination of the role of working memory in simple verbal inferences. They have found that when subjects are asked to hold in mind a string of digits, their performance in reasoning tasks is adversely affected in comparison with a control group that had no such load on memory (see Baddeley and Hitch, 1974; Hitch and Baddeley, 1976). It is therefore plausible to suppose that the effect of having to construct a greater number of models has its locus in working memory.

Individual differences in reasoning ability

What causes people to differ in their ability to make inferences? That they do differ is, of course, evident from the longstanding use of syllogisms in tests of intelligence. Yet no one knows for certain what aspects of mental processing make one person a good reasoner and another a poor reasoner. Whatever the general merit of investigating 'individual differences' by way of mental tests, their use is unfortunately of little value in the study of reasoning. The data they yield are, as I argued in Chapter 4, too gross to elucidate differences in mental processes from one individual to another.

The theory of mental models offers an explanatory framework that helps both to make sense of differences in reasoning ability and to go beyond a merely 'actuarial' account of mental processes. It specifies the separate components underlying inferences and places constraints on the possible differences between individuals. The theory assumes that inferences depend on three component skills: (1) the ability to form an integrated model of the premises; (2) the appreciation that an inference is sound only
if there are no counter-examples to it, together with the capacity to put this principle into practice, and (3) the ability to put into words the common characteristics of a set of mental models. Bruno Bara and I have explored the differences in the protocols of subjects carrying out syllogistic inferences, and I shall summarize the outcome of our research.

The only way to convey the 'feel' of the results is to present data from individual subjects. Table 5.5 gives the percentages of valid conclusions drawn by twenty American students, and Table 5.6 gives the percentages for twenty Italian students (the results are from Experiments 1 and 2 reported earlier). Both tables show the percentages as a function of the number of mental models that have to be constructed to draw a valid conclusion. Fortuitously, the two samples of subjects differed in their inferential ability. The most striking overall pattern in the two samples of data is the decline in performance as the number of models to be constructed increases: thirty-three out of the forty subjects conform precisely to this predicted trend, and of the remainder only two showed competence.

Table 5.5 The percentages of correct conclusions drawn by twenty American subjects in an experiment on syllogistic reasoning. The percentages are shown as a function of the number of models to be constructed in order to draw a valid conclusion (Johnson-Laird and Bara, 1982)

<table>
<thead>
<tr>
<th>Subjects</th>
<th>One model (n = 10)</th>
<th>Two models (n = 5)</th>
<th>Three models (n = 12)</th>
<th>Overall (n = 27)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>100</td>
<td>85</td>
<td>63</td>
<td>56</td>
</tr>
<tr>
<td>2</td>
<td>90</td>
<td>63</td>
<td>42</td>
<td>48</td>
</tr>
<tr>
<td>3</td>
<td>60</td>
<td>42</td>
<td>25</td>
<td>37</td>
</tr>
<tr>
<td>4</td>
<td>80</td>
<td>63</td>
<td>25</td>
<td>56</td>
</tr>
<tr>
<td>5</td>
<td>80</td>
<td>63</td>
<td>25</td>
<td>56</td>
</tr>
<tr>
<td>6</td>
<td>90</td>
<td>63</td>
<td>25</td>
<td>56</td>
</tr>
<tr>
<td>7</td>
<td>80</td>
<td>63</td>
<td>25</td>
<td>56</td>
</tr>
<tr>
<td>8</td>
<td>90</td>
<td>63</td>
<td>25</td>
<td>56</td>
</tr>
<tr>
<td>9</td>
<td>80</td>
<td>63</td>
<td>25</td>
<td>56</td>
</tr>
<tr>
<td>10</td>
<td>90</td>
<td>63</td>
<td>25</td>
<td>56</td>
</tr>
<tr>
<td>11</td>
<td>80</td>
<td>63</td>
<td>25</td>
<td>56</td>
</tr>
<tr>
<td>12</td>
<td>90</td>
<td>63</td>
<td>25</td>
<td>56</td>
</tr>
<tr>
<td>13</td>
<td>100</td>
<td>63</td>
<td>25</td>
<td>56</td>
</tr>
<tr>
<td>14</td>
<td>100</td>
<td>63</td>
<td>25</td>
<td>56</td>
</tr>
<tr>
<td>15</td>
<td>100</td>
<td>63</td>
<td>25</td>
<td>56</td>
</tr>
<tr>
<td>16</td>
<td>80</td>
<td>63</td>
<td>25</td>
<td>56</td>
</tr>
<tr>
<td>17</td>
<td>90</td>
<td>63</td>
<td>25</td>
<td>56</td>
</tr>
<tr>
<td>18</td>
<td>80</td>
<td>63</td>
<td>25</td>
<td>56</td>
</tr>
<tr>
<td>19</td>
<td>80</td>
<td>63</td>
<td>25</td>
<td>56</td>
</tr>
<tr>
<td>20</td>
<td>80</td>
<td>63</td>
<td>25</td>
<td>56</td>
</tr>
<tr>
<td>Overall%</td>
<td>92</td>
<td>63</td>
<td>25</td>
<td>56</td>
</tr>
</tbody>
</table>

contrary to the predictions - subjects 1 and 2 in the Italian sample were able to cope with three-model problems, but not two-model problems. I will now examine the individual data in the light of the three main components of the theory.

The process of forming an integrated mental model of premises is nothing more than the proper comprehension of discourse: it is required in order to grasp the full impact of what a speaker has to say. The ability to carry it out should be common to all native speakers of a language, and, since it and its complementary skill of putting models into words suffice for competency with syllogisms requiring only one model, it is hardly surprising that the subjects were almost universally able to cope with these syllogisms. Every single subject performed more accurately with them than with any other sort (for those who relish significance levels, the chance probability of such a result is less than one in a billion). The main difficulty in constructing an integrated model is that a representation of one premise...
Mental models

must be held in working memory while information from a representation of the other premise is combined with it. The two subjects (19 and 20) in the Italian experiment who failed to do better than chance with one model were quite unable to cope, as one would expect, when the premises required more than one model to be constructed. Likewise, the figural arrangement of terms had a striking effect on their performance: they could only form a model for premises of the A – B, B – C and B – A, C – B figures. With premises in the other figures, which require interpretations to be switched round, they either indicated erroneously that there was no valid conclusion or else forgot one of the end terms and mistakenly replaced it with a middle term so as to form a conclusion that was blatantly inconsistent with one of the premises. Their tendency to assert that there was no valid conclusion if (and only if) the figures required interpretations to be switched round gave rise to a spuriously good performance with syllogisms that have no interesting conclusions in these figures.

Only where a valid inference depends on constructing alternative models of the premises are genuine differences in inferential ability to be observed. A reasoner must appreciate the need to construct and to evaluate alternative models, and must be able to carry out this procedure within the processing limitations of working memory. An important general point is that the subjects' performance with valid syllogisms and their performance with invalid syllogisms is not reliably correlated ($\tau = 0.11$, $p > 0.2$, for the American subjects; $\tau = -0.13$, $p > 0.2$, for the Italian subjects). The lack of a correlation arises from the responses to those premises for which more than one model can be constructed, which include, of course, all the problems for which there is no valid conclusion. A few subjects seem not to appreciate the need to consider alternatives. The hallmark of their performance is a string of erroneous conclusions and a reluctance to respond that there is no valid conclusion. None of the Americans fell into this category: even subject 20 got 32% of the problems with no valid conclusion correct. There were, however, three Italian subjects (13, 14 and 15) who performed poorly with premises requiring more than one model and who responded correctly to the problems lacking a valid conclusion on less than 20% of occasions. Subject 15, in particular, got only two of these thirty-seven problems correct, otherwise drawing a conclusion based on a single model of them. Other subjects perceive the need to construct alternatives, and are able to do so, but are wholly incapable of assessing them. The hallmark of their performance is a tendency to respond 'No valid conclusion', whether or not it is justified. They accordingly perform sp uriously well with premises that do not yield a conclusion, but fall down badly with premises that require assessing more than one model to yield a valid conclusion. Any subject who performs better with invalid syllogisms than with valid syllogisms is showing symptoms of this syndrome. Among the Americans, four subjects (11, 14, 17 and 18) showed marked signs of it: e.g. subject 17 got 89% of the invalid problems correct, but also tended to respond 'No valid conclusion' to valid problems which required more than one model to be constructed. The two most striking instances of the syndrome among the Italians were subject 16 (81% of the invalid syllogisms correct) and subject 17 (62% of the invalid syllogisms correct).

Most subjects are able to construct some alternative models, but from time to time they fall down in assessing what, if anything, they have in common. They are particularly prone to error in those figures that require interpretations to be switched round, failing to detect either that a putative conclusion is violated by one alternative model or else that there is a conclusion common to all the alternatives. It is noteworthy that only one out of the forty subjects in the two experiments showed any competence in dealing with the most difficult syllogisms of all, namely, those with three models where the conclusion runs counter to figure.

There are several other differences in performance between the subjects, including their susceptibility to figural effects, which I shall not analyse here. My point has been to establish that the theory of mental models at least provides a framework suitable for describing individual differences, and even suggests some explanations for them. Indeed, Jane Oakhill and I have recently confirmed in an unpublished study that there is a reliable correlation ($p = 0.7$) between a simple measure of the processing capacity of working memory – the number of pairs of letters (e.g. 1B) that can be converted in a fixed interval of time into those that are two places later in the alphabet (KD) – and accuracy in syllogistic reasoning. Of course, there are other general personal characteristics, such as impulsivity, that are likely to affect reasoning ability just as they affect the performance of any other intellectual task.

**Children's ability to reason syllogistically**

Such is Piaget's influence on the study of intellectual development that the majority of psychologists, whether or not they subscribe to his theories, probably believe that children learn to reason formally by the age of 11 to 12 years – the age of Piaget's putative stage of 'formal operations'. There
have been many studies which establish that children are able to make
deductions at a much earlier age than is dreamt of in Piagetian lore (see,
for example, Mehler and Bever, 1967; Bryant and Trabasso, 1971; Donald-
son, 1978). But what is equally important is that there are varieties of
deduction that are mastered only after the age of twelve. The best evidence
comes from the study of syllogisms.

Debbie Bull and I designed an experiment to investigate the ability of
two groups of intelligent children (nineteen 9- to 10-year-olds and nineteen
11- to 12-year-olds) to draw their own conclusions from syllogistic premises.
It was not possible to use all sixty-four pairs of premises – the task would
have been too gruelling – but we carefully selected twenty pairs (of which
sixteen yielded valid conclusions) in order to examine the children’s general
competence. The experiment showed that there was only a slight difference
between the abilities of the two groups, that the children were just as
susceptible to the effects of figure as adults, but that, unlike most adults,
they could only draw valid conclusions that depended on constructing one
mental model. The older group responded correctly on 58% of the one-
model problems – there were eight such problems and a child had to get
four of them right to perform reliably better than chance. Not one of these
older subjects drew a correct conclusion to the two-model problem or to
the six three-model problems. The younger group responded correctly to
41% of the one-model problems, three subjects (16%) responded correctly
to the two-model problem, and not a single subject responded correctly
to the three-model problems. In short, some of the children were competent
with one-model problems, three of them might have been competent with
the two-model problem, but none of them was competent with three-model
problems. The syllogism that elicited the best performance from both
groups was of the form:

All the A are B
All the B are C
∴ All the A are C.

Doubtless, Aristotle would have predicted this result, because the syllogism
is in the form that he considered to be perfect.

Granted that the children taking part in the experiment are intelligent
– and the teachers selected the brightest of their children – then they
should at a later stage develop a logical ability on a par with the adults
that we have tested: they should improve significantly in their ability to
reason syllogistically.

Educational applications: how to improve reasoning ability

Educationalists have developed a variety of methods designed to improve
the ability to reason. They include the pedagogical use of stories illustrating
logical principles (Lipman and Sharp, 1978), the deployment of special
reasoning problems (Feuerstein, Hoffman, and Miller, 1980), and various
courses on thinking and problem solving (e.g., Whimbey and Lochhead,
1980). Psychologists have become increasingly involved in such matters,
especially since the start of the project to raise the intelligence of the entire
population of Venezuela (the international newsletter Human Intelligence
has published several reports on this project, which includes work carried
out by researchers at Harvard University, Bolt Beranek and Newman, Inc.,
and many other research organizations). Applied psychologists need to
devise an economical technique for assessing the strengths and weaknesses
of an individual’s inferential skills, and they need to implement effective
remedial procedures that will overcome the various deficiencies underlying
poor performance. The work that my colleagues and I have carried out
has two main implications. First, it is important to diagnose and to
distinguish between weaknesses in reasoning ability that result from the
following factors:

1. Cultural or personal characteristics that underlie apparently poor
performance, but that have nothing whatsoever to do with basic intellectual

2. Linguistic impairments that make it difficult to understand and to
remember verbal premises, or to put into words conclusions that have
been drawn.

3. Failures to appreciate the fundamental principle of valid inference,
or to understand that a reasoning test calls for conclusions that follow of
necessity.

4. An inability to construct integrated representations of the premises,
or to evaluate alternative representations, as a result of a limited working
memory.

Such deficiencies are readily detectable if subjects are asked to state what
conclusions follow of necessity from the sixty-four different pairs of syllo-
logistic premises. The premises should each be presented with a sensible
everyday content that does not predispose subjects to any particular
conclusion on the basis of general knowledge. If the test is followed up
by a ‘debrieving’ session in which the subjects are asked to explain the
reasons for their fallacious conclusions, it is relatively easy to identify the causes of their particular problems.

Second, our work suggests that the most common cause of difficulty in reasoning for individuals living in a literate society is the limited processing capabilities of working memory. Its effects have been apparent in every subject that we have tested (see, e.g., Tables 5.5 and 5.6). It must be emphasized, however, that there is a spontaneous improvement in reasoning ability simply as a consequence of practice (with no feedback). The subjects whose data are reported in Table 5.5 were tested again one week later. They were given no forewarning that they would be retested, but their overall performance increased by 10%, and nineteen out of the twenty subjects returned an improved score. One striking differential effect of practice occurred with the valid conclusions drawn from premises in the most difficult figure, B - A, B - C. Here, there was an overall improvement of 20%, and half of it was due to a decline in erroneous responses to the effect that there was no valid conclusion. The result of practice must in part be to increase the efficiency of the encoding operations of working memory, enabling subjects to switch round the interpretation of a premise. The B - A, B - C figure is difficult according to the mental model theory, precisely because it is always necessary to switch round the interpretation of at least one premise. Subjects who earlier were unable either to form alternative integrated models, or to assess their implications, improved with practice. Experience with the task may also produce a growing awareness of general principles governing the logical properties of the problems. Some subjects may notice, for example, that two negative conclusions never yield a valid conclusion and in this way they may begin to perform like logicians and no longer need to construct mental models.

Some commentators on our work have suggested that diagrams of mental models might serve a useful pedagogical function in teaching people the principles of deduction. Although the prospect is appealing, it may be dangerous. Whenever I have presented a reasoning problem informally, I have noticed the difficulties that people get themselves into if they use Euler circles. The problem is that there is no simple algorithm for using them that one can learn as one learns, say, the algorithm for long multiplication. Merely drawing circles does not guarantee that all their possible combinations will be considered exhaustively. The same problem applies to the notation of mental models - if there were a simple algorithm that guaranteed an exhaustive search, doubtless most of us would have mastered it when we first learnt to reason. Educators are probably better advised
to ensure that their students understand the fundamental principle of inference and get plenty of opportunities to put it into practice - at least until someone should chance upon an effective way of increasing the processing capacity of working memory.

Conclusions

This chapter has outlined a theory of syllogistic inference based on the assumption that reasoners construct integrated mental models of the premises. These models have an important structural property deriving from a constraint on the set of possible mental models: a natural mental model of discourse has a structure that corresponds directly to the structure of the state of affairs that the discourse describes. The sophisticated notations of Euler circles and Venn diagrams lack this property, and consequently they are not natural mental models. For example, a premise such as:

All the artists are beekeepers

describes a state of affairs in which one finite set of individuals is mapped into another. A natural mental model likewise contains one finite set of individual tokens mapped into another. Neither Euler circles nor Venn diagrams, however, contain finite sets of individual tokens: they map finite sets of individuals into infinities of points. Because any syllogistic premise can be represented in a single mental model of the present sort, the theory avoids the combinatorial problems that bedevil Euler circles, and similarly the maximal load on working memory (three different models) is considerably less than would be required for Venn diagrams (eight different contingencies). The theory has been corroborated experimentally and the results that have been presented establish its descriptive adequacy according to the criteria introduced in Chapter 4; it accounts for the relative difficulty of different syllogisms and for the systematic errors and biases in performance (criterion 1); it describes the characteristic patterns of individual differences (criterion 2); and it also has some implications for pedagogy (criterion 7). The theory extends to other sorts of deductive inference, and I propose to evaluate the explanatory adequacy of the general theory rather than the specific instance of it presented here.
Inference and mental models

At the end of a lecture on ethics, Epictetus, the Stoic philosopher, recommended the study of logic to his audience because it was useful. One of his listeners was unconvinced, and asked: ‘Sir, would you demonstrate the usefulness of the study of logic?’ Epictetus smiled and replied: ‘That is my point. How could you, without the study of logic, test whether my demonstration would be valid or not?’

The late Yehoshua Bar-Hillel, who recounts this story, comments that if the audience had signed up to take the course in logic that Epictetus announced, then they must surely have been very disappointed. Epictetus had shown merely that there was a need for a theory that would make it possible to test the validity of arguments in ordinary language, not that he possessed such a theory. Indeed, as Bar-Hillel (1970) emphasizes, logic until very recently was insufficiently powerful to cope with natural language.

There is another, very different, moral to the anecdote about Epictetus. His audience may or may not have been disappointed by his course, but doubtless they were capable of thinking rationally in any case. Valid inferences were made long before the invention of logic; and they can be made without relying, consciously or unconsciously, on rules of inference. This point will be clarified in assessing the general theory that inferences are based on mental models. In the previous chapter, a special case of the theory concerning syllogisms was shown to be descriptively adequate. In this chapter, I shall describe the general theory of inference, which incorporates the machinery for the ‘test case’ of syllogisms, and evaluate its logical and psychological properties in the light of the explanatory criteria that any adequate theory of reasoning should satisfy. In particular, I shall consider the variety of inferences with which it deals – implicit and explicit – how children master them, and how validity is captured within the theory without recourse to rules of inference.

 Implicit inferences

There is an important distinction between two sorts of inference that occur in daily life. On the one hand, the inferences that I have so far considered mostly require a conscious and cold-blooded effort. You must make a voluntary decision to try to make them. They may take time and they are at the forefront of your awareness: they are explicit. On the other hand, the inferences that underlie the more mundane processes of intuitive judgment and the comprehension of discourse tend to be rapid, effortless, and outside conscious awareness: they are implicit. Suppose, for example, you were to read in the paper:

There was a fault in the signalling circuit. The crash led to the deaths of ten passengers . . .

then you might well infer that the passengers were killed in the crash. The text does not make this assertion, and it might even continue:

because they were arrested after the accident, and subsequently shot as spies.

Plainly, you jumped to a conclusion based partly on the content of the passage and partly on your general knowledge. You make such inferences automatically, almost involuntarily, and often without being aware of what you are doing. Since a valid inference is one for which, if the premises are true, the conclusion must be true, an important feature of these inferences is that they are usually invalid.

There are many sorts of inference, but any adequate psychological theory of reasoning must recognize the distinction between implicit and explicit inferences. When theorists have tried to formulate theories of thinking, however, there has been a strong tendency for them to concentrate on explicit rather than implicit reasoning. In the seventeenth century, Pascal (trans. Krailsheimer, 1966, p. 211) observed:

It is rare for mathematicians to be intuitive or the intuitive to be mathematicians, because mathematicians try to treat intuitive matters mathematically, and make themselves ridiculous, by trying to begin with definitions followed by principles, which is not the way to proceed in this kind of reasoning. It is not that the mind does not do this, but it does so tacitly, naturally and artlessly, for it is beyond any man to express it and given to very few even to apprehend it.

The point is well taken: psychologists had largely overlooked implicit inferences until attempts to program computers to understand discourse revealed their ubiquity. The credit for this discovery must go to Charniak.
Minsky, Schank, and Winograd, and to their colleagues on this side of the Atlantic, Isard, Longuet-Higgins, and Wilks.

Without an ability to make implicit inferences, written and spoken discourse would be beyond anyone's competence. In order to understand the following passage, it is necessary to make a variety of inferences:

The pilot put the plane into a stall just before landing on the strip. He just got it out of it in time. Wasn't he lucky?

Every word in the first sentence, apart perhaps from the articles and prepositions, is ambiguous, and the appropriate meanings can be recovered only by making implicit inferences from linguistic context and general knowledge. To make sense of the second sentence, a number of inferences have to be made to determine the referents of the pronouns: the first *it* refers to the plane, and the second, to the stall. The third sentence is not to be taken as a question, though it is interrogatory in form: an inference from the context establishes that it has the force of an assertion. At the point at which most of these inferences are made they can seldom be securely established: they are plausible conjectures rather than valid deductions. Many psychologists are accordingly inclined to suppose that they must depend on the computation of probabilities. However, there is no need to suppose that individuals compute probabilities in determining, say, that a pronoun refers to one entity rather than another. The mechanism is more likely to consist of a device that constructs a single mental model on the basis of the discourse, its context, and background knowledge. Such knowledge is embodied in the model by default, that is, it is maintained in the model provided that there is no evidence to the contrary. No attempt is made to search for an alternative model unless such evidence arises. It is for this reason that the process can be very rapid; it becomes as automatic as any other cognitive skill that calls for no more than a single mental representation at any one time. And it is also for this reason that implicit inferences lack the guarantee, the mental imprimatur, associated with explicit deductions. Hence, the fundamental distinction between the two types of inference is whether or not there is a deliberate search for alternative models of the discourse.

The acquisition of implicit inferential ability

If the present account of implicit inferences is correct, then children must acquire the ability to make such inferences in order to understand discourse.

This hypothesis has been borne out by a number of experimental studies. Til Wykes, a former student of mine, showed that young children (about 4 years old) have considerable difficulty in correctly acting out with glove puppets such pairs of sentences as:

Susan needed Mary's pencil.
She gave it to her.

The task is much easier for them if gender can be used as a cue:

Susan needed John's pencil.
He gave it to her.

In general, the greater the number of pronouns in a sentence, the harder it is for young children to understand it properly. They appear to adopt a syntactically-based procedure for assigning referents to pronouns rather than an inferential one. They assume that a pronoun is co-referential with the subject of the previous clause (see Wykes, 1981). In a further study, we discovered that children are poor at making implicit inferences to work out the meaning of such sentences as 'The Smiths saw the Rocky Mountains flying to California' (Wykes and Johnson-Laird, 1977). Similarly, children presented with a sentence such as:

The man stirred his cup of tea
tend not to infer spontaneously that the man used a spoon to stir his tea. In all these cases, it was clear from control studies that the children are able to make the relevant inferences. The point is that they do not do so as a matter of course in understanding discourse.

The ability to make implicit inferences is equally important, of course, for reading. Jane Oakhill, another former student, has shown that an important distinction between excellent and average readers is precisely this inferential ability. In one study, Oakhill (1982) gave a sample of 168 children (aged 7 to 8 years) a variety of vocabulary and reading tests. She was then able to select two groups matched on vocabulary and phonics skills, but differing considerably in their ability to understand what they read. The two groups of children were then tested in an experiment on their ability to make inferences when listening to simple stories. Each story consisted of three sentences, such as:

The car crashed into the bus.
The bus was near the crossroads.
The car skidded on the ice.

After the children had heard eight such stories, their memory for them
was examined in a recognition test. A child who has built up an integrated mental model on the basis of implicit inferences is likely to assume that the sentence:

The car was near the crossroads

had originally occurred in the story. Given the nature of the original events, this inference is extremely plausible. But a sentence such as:

The bus skidded on the ice

is much less plausibly inferred, since there is no reason to make this inference in building a model of the events in the story. The results of the recognition test using such sentences showed, as expected, that the good readers tended to make more errors based on plausible inferences than did the average readers. Good readers, however, performed better than average readers in recognizing the original sentences from the stories and in rejecting the implausible inferences. It seems that good readers are more likely than average readers to make implicit inferences in order to build up a mental model of a story. This study obviously tells us nothing about causal direction: good readers may be good because they spontaneously make inferences, or they may make such inferences because they are good readers as a result of other factors. In a series of additional experiments, however, Oakhill failed to find any other major difference in the abilities of her two groups of readers.

One addendum to this work is worth noting: The procedure is based on one devised by Paris and his colleagues (e.g., Paris and Carter, 1973) though these investigators were not concerned with differences in reading ability. Their studies have been criticized on the grounds that the sentences used in the memory tests allowed children to detect the new sentences, which had not occurred in the original stories, solely on the grounds that they contained new words that also had not occurred in the original stories. The materials used by Oakhill, as the example above shows, were carefully selected so as to obviate this criticism.

Reasoning without rules of inference

Logical thinking in daily life is most likely to occur in explicit inferences. From a semantic standpoint, the natural way to think of logic is as a set of procedures for establishing the validity of a given inference, that is, as a method for showing systematically that there is no interpretation of the premises that is consistent with the denial of the conclusion. From a formal or syntactic standpoint, a logic contains a set of rules of inference, or inferential schemata, that allow conclusions to be formally derived from premises. Explicit inferences based on mental models, however, do not need to make use of rules of inference, or any such formal machinery, and in this sense it is not necessary to postulate a logic in the mind. This claim, as I know from the reaction of audiences to whom it has been addressed, is both hard to understand and hard to believe – it is viewed as almost on a par with the Pelagian heresy in some quarters. That the doctrine of original sin may be mistaken is hardly a new idea, but perhaps there is something novel in the notion that the doctrine of mental logic may be wrong – the major tradition in psychology, culminating in the work of Jean Piaget, has always firmly implanted formal logic in the mind.

The crux of the matter is that a system of inference may perform in an entirely logical way even though it does not employ rules of inference, inferential schemata, or any other sort of machinery corresponding to a logical calculus. The rest of the argument is simple once this point is grasped, so let me labour it awhile.

I am going to describe a hypothetical procedure for making syllogistic inferences from affirmative premises which is related to the theory presented in the previous chapter, but which is designed to illustrate in the clearest and most elementary way the feasibility of reasoning without formal rules of inference. The first step in the procedure is to construct a finite model that satisfies the first premise and a finite model that satisfies the second, taking care to use the same number of items to represent the middle term in both models. Next, the procedure forms an integrated model by joining the two in virtue of the items representing the middle term. It then formulates a conclusion that is true of the resulting model and that interrelates the two end terms. So far, of course, it has not employed any logic: the conclusion may well be blatantly invalid. For example, given the premises:

All the A are B
All the C are B

the procedure forms the model:

\[
\begin{align*}
a &= b = c \\
(b) \\
(b)
\end{align*}
\]
and draws the erroneous conclusion:

All the A are C.

However, it now begins an exhaustive series of tests whose outcome, if any, will be a valid conclusion. It selects an end item in the model at random. If there is an identity between that end item and a middle item, it destroys it and seeks to form a new identity between the end item and a middle item hitherto unrelated to the class of end items. In the example, the procedure might select the second token corresponding to C, destroy its present identity and establish a new one:

\[ a = b = c \]
\[ a = b \]
\[ (b) = c \]
\[ (b) \]

If the end item selected at random is not initially linked by an identity to a middle item—an eventuality that could arise with other premises—then the procedure seeks to establish such a relation. In either case, it then checks whether the resulting model is still consistent with the premises, and, if it is, whether the model is consistent with the current conclusion. Where the model is consistent with the premises but not consistent with the conclusion, as in the example, that conclusion is rejected as invalid.

The procedure formulates (if possible) a new conclusion covering the current model and all the previous models consistent with the premises. It then samples a new end item at random and carries out the whole process again. Since the model is finite, then, provided that the procedure records the selections that it has made, there will eventually be no more items to select. If there is any conclusion still remaining at that point, then it is valid.

This procedure is extremely wasteful even for the restricted set of inferences with which it can cope. It constructs many models that violate the truth conditions of the premises, and which are therefore useless. Yet, although it works randomly, it does form a complete search. It embodies the general semantic principle that lies behind all logics, though not explicitly formulated within any of them: an inference is valid if and only if there is no interpretation of the premises that falsifies the conclusion. However, the procedure does not employ any formal logical apparatus.

That is why it has to search for falsifications at random and wastes most of its time looking at models that are not even relevant to the question of validity.

What about the machinery for searching for end items, destroying identities, and testing truth conditions—does it perhaps incorporate a formal logic? Once again, the answer is negative. All that is required is the ability to construct models, to search for entities in them, and to generate descriptions of them—in short, the basic computational power described in Chapter 1 in terms of Turing machines and recursive functions. Such functions do not in themselves constitute a logic: they can be used to model either logical or illogical processes. They can even be used to develop a theory of human inference that allows for both rationality and errors.

The general theory of explicit inference based on mental models assumes that human reasoners can construct finite models of premises, formulate putative conclusions on the basis of them, and search for models of the premises that are counter-examples to such conclusions. If the search is systematic and exhaustive, then the conclusion is valid. But human reasoners often fail to be rational. Their limited working memories constrain their performance. They lack the guidelines for systematic searches for counter-examples; they lack secure principles for deriving conclusions; they lack a logic. Since even the most intelligent individuals have difficulty with certain syllogisms, and are aware of it, they have an obvious motivation to try to externalize and to systematize the search for counter-examples.

Hence, the theory is compatible with the development of logic as an intellectual tool.

When Aristotle invented logic, his method was to determine which pairs of syllogistic premises yielded valid conclusions (see Kneale and Kneale, 1962). An inference of the form:

Every man is an animal
No stone is a man
\[ \therefore \] No stone is an animal

certainly yields a true conclusion. In order to determine whether inferences of this form were valid, Aristotle changed the content of the premises whilst holding their form constant, e.g.:

Every man is an animal
No horse is a man
\[ \therefore \] No horse is an animal.

The conclusion is manifestly false, but the inference is identical in form to the previous example. Since the form can lead to false conclusions from true premises, it must be invalid. Instead of searching for interpretations
of premises that are counter-examples to conclusions, Aristotle held form constant and searched for premises with a content that was incompatible with the corresponding conclusion. He used his semantic intuitions to determine the set of valid syllogisms, and then he developed a logic—a set of principles for deriving validity.

The power of mental models

A major advantage of natural mental models over other, more sophisticated, forms of representation such as Euler circles, Venn diagrams, and even the first-order predicate calculus, is that they can represent the content of any sentences for which the truth conditions are known. I shall illustrate this point by considering inferences based on relations and on different sorts of quantifiers.

Relational expressions such as greater than, father of, next to, equals, loves, have a number of different logical properties. A relation such as greater than is transitive because it permits an inference of the form: if x is greater than y, and y is greater than z, then x is greater than z. In more general terms, the transitivity of a relation R guarantees the validity of an inference of the form:

\[ xRy \text{ and } yRz \therefore xRz \]

A relation such as father of is intransitive because it leads to the negation of such a conclusion:

\[ xRy \text{ and } yRz \therefore \neg (xRz) \]

A relation such as next to is neither transitive nor intransitive. A second set of properties concerns the symmetry of relations. A relation such as next to is symmetric because if x is next to y, then it follows that y is next to x. A relation such as greater than is asymmetric because if x is greater than y then it follows that y is not greater than x. A relation such as not greater than is neither symmetric nor asymmetric. A third set of properties concerns the reflexivity of relations. A relation such as equals is reflexive because for any entity x, it is always the case that x equals x. A relation such as taller than is irreflexive because for any entity x, it is never the case that x is taller than x. A relation such as loves is neither reflexive nor irreflexive since it fails to guarantee either conclusion—some individuals love themselves, and some do not.

Psychologists have devoted most of their study of relations to the topic of transitivity. An argument of the sort:

Alice is taller than Betty
Betty is taller than Carol
\[ \therefore \text{Alice is taller than Carol} \]

is treated by logicians as technically invalid, since it lacks a premise to the effect that taller than is transitive. Advocates of mental logic are accordingly forced to argue either that there is a general schema of transitivity:

For any x, y, and z, if xRy and yRz then xRz

to which particular relations such as taller than become linked, or else that specific schemata are acquired for each transitive relation in the form of specific rules or ‘meaning postulates’:

For any x, y, and z, if x is taller than y, and y is taller than z, then x is taller than z.

Granted a knowledge of the principle of transitivity, children might readily learn to which relations it applies. The problem, however, is to explain the acquisition of the principle itself. Once the idea of mental logic is given up, that problem is reduced to one of accounting for how the meanings of relational terms are learned. This problem, as I shall show, is more tractable.

The relational expressions more than, taller than, bigger than, kinder than and their cognates are all transitive, asymmetric, and irreflexive. It follows that the differences in meaning between them cannot be distinguished merely on the basis of their logical properties. An account of their semantics is needed in any case, since a complete specification of the meaning of a relational assertion such as:

There are more a’s than b’s

must specify its truth conditions. If the domain of discourse consists of sets of entities, then a semantics for this assertion can be specified in terms of a one-to-one mapping. The assertion is true if and only if each b can be put into a one-to-one correspondence with an a, but there remain outstanding a’s not within the mapping. If you have acquired this semantics for more than, then you can interpret the assertion ‘There are more a’s than b’s’ by forming a mental model containing some representative number.
of members of the two classes with the following mapping between them:
\[
\begin{align*}
a &\rightarrow b \\
a &\rightarrow b \\
a &\rightarrow b \\
a
\end{align*}
\]
Likewise, you can interpret the assertion ‘There are more b’s than c’s in the same manner:
\[
\begin{align*}
b &\rightarrow c \\
b &\rightarrow c \\
b
\end{align*}
\]
So, if you combine your two interpretations, taking pains to ensure that there were the same number of b’s in both, you will form the following unified representation:
\[
\begin{align*}
a &\rightarrow b &\rightarrow c \\
a &\rightarrow b &\rightarrow c \\
a &\rightarrow b \\
a
\end{align*}
\]
You could also readily ascertain that the truth conditions of the assertion ‘There are more a’s than c’s’ are satisfied by your mental representation. Hence, the semantics of the relation ensure that whenever there are more a’s than b’s and more b’s than c’s then there are more a’s than c’s. You can make the transitive inference by building a mental model of this sort even though you have neither a meaning postulate for transitivity nor an internal logic of relations. In the same way, a mental model establishes the asymmetry and irreflexiveness of more than: if there are more a’s than b’s then the converse cannot hold; and there can never be more a’s than a’s. The logical properties of a relation emerge naturally from its meaning.

The same principle of emergent logical properties applies to other relational terms. In the case of length, for example, the truth conditions of ‘a is longer than b’ can be specified either in terms of the cardinality of the units of measurement, i.e., the number of units for a are more than the number for b, or in terms of a direct comparison between the major axes of a and b. The assertion ‘a is longer than b’ is true if and only if the major axis of a is coextensive with a proper part of the major axis of b. If you have acquired a semantics, then you can construct a mental model in which a and b are lined up so that b’s axis is coextensive with only part of a’s axis. There are obvious problems about the direct encoding of real numbers, since the continuum is infinitely and uncountably dense and could not be directly represented in a finitary device such as the human brain. However, it is plausible to suppose that lengths are mentally represented by a digital approximation. This principle would permit an alternative representation of ‘a is longer than b’ in which a and b are located in appropriate positions on a single ‘dimension’ corresponding to length—a format that has been postulated by several authors (De Soto, London, and Handel, 1965; Hutterlocher, 1968), and that could apply to such relations as ‘better than’, ‘prettier than’, ‘kinder than’. A combined representation of separate premises, whether in terms of axes or a unitary dimension of length, would readily yield a transitive conclusion.

Piagetian theorists have assumed that there is a transitivity schema and that it derives from the notion of serial order: ‘transitivity is an ordinal qualitative property’ (Youniss, 1975, p. 234). Although it is true that serial orders give rise to transitivity, there are a number of transitive relations that do not depend on orderings, e.g., ‘is included in’, ‘is an ancestor of’, ‘is identical to’. Their transitivity follows from their semantics, and can be readily grasped from an appropriate mental model: for instance, a model in which a is included in b and b is included in c readily yields the conclusion that a is included in c. It is accordingly a mistake to try to reduce the representation of all transitive relations to a system based on an underlying ordinal dimension.

Turning to quantifiers, mental models cope naturally with the interpretations of specific numbers and such terms as most, many, several, and a few. Janet Fodor (1982) has argued for a very similar form of mental representation in order to distinguish between the meaning of such quantifiers as every and all. Mental models can certainly provide the necessary interpretations for inferences that hinge on relative sizes, e.g.,

All fascists are authoritarians.
Most authoritarians are dogmatic.

These premises are representable by the following type of tableau:
\[
\begin{align*}
f &\rightarrow a \\
f &\rightarrow a \\
f &\rightarrow a \\
f &\rightarrow a \\
f &\rightarrow a \\
f &\rightarrow a \\
f &\rightarrow a \\
f &\rightarrow a \\
f &\rightarrow a \\
f &\rightarrow a \\
f &\rightarrow a \\
\end{align*}
\]
which suggests the conclusion:

Most fascists are dogmatic.
MENTAL MODELS

People readily make these sorts of inference even though their validity is problematical. It seems to depend on the relative sizes of sets. The following premises are superficially of the same form as those in the example above, but they are unlikely to elicit a comparable conclusion:

All archbishops are tories.
Most tories are middle class.

Even if these premises were true, one knows that the number of archbishops is very small in comparison to the number of tories. Hence, the premises are entirely compatible with the model:

\[
\begin{align*}
  a &= t \\
  t &= m
\end{align*}
\]

In cases of this sort, there is a positive advantage in using mental models rather than a mental logic to make inferences. A logic will either sanction an inference or not (though the particular status of the inference may not be obvious). A system of mental models more accurately reflects the uncertainty of inferences that depend on proportions: a conclusion of debatable status is forthcoming, since the search for refutations can be guided by general knowledge.

The distinction between class-membership and class-inclusion is represented straightforwardly in a mental model. The premise:

Jock is a Scotsman

calls for the following sort of model:

\[
\begin{align*}
  s \\
  j &= s \\
  s \\
  s
\end{align*}
\]

where \( j \) represents Jock, and the \( s \)'s represent the complete class of Scotsmen, since there is no optional element signifying that the term is not distributed. The premise:

Scotsmen are widely scattered around the world

asserts that the class of Scotsmen is included in the class of classes whose members are widely scattered around the world. This assertion can be represented by a model in which the class of Scotsmen as a whole is identified with a class of individuals widely scattered around the world:

\[
\begin{align*}
  s_1 &= w \\
  s_2 &= w \\
  s_3 &= w \\
  s_4 &= w
\end{align*}
\]

where the brace indicates that the set of individuals that it embraces is identified with a member of the class of \( w \)'s, classes of entities whose members are widely scattered. The two models can be combined:

\[
\begin{align*}
  s_1 &= w \\
  j &= s \\
  s_2 &= w \\
  s_3 &= w \\
  s_4 &= w
\end{align*}
\]

But since it is the class of Scotsmen as a whole that is identified with a class of entities scattered around the world, there is no way of forming an identity that yields the conclusion that Jock is widely scattered around the world. Class-inclusion, however, is represented in a way that automatically yields a transitive inference. The premises:

Jock is a Scotsman
Scotsmen are human beings

are representable as:

\[
\begin{align*}
  s &= h \\
  s &= h \\
  j &= s = h \\
  s &= h
\end{align*}
\]

and a direct link is established between \( j \) (Jock) and \( h \) (a human being). It follows that Jock is a human being.

Mental models also accommodate assertions containing multiple quantifiers. A statement such as 'Some professors attended all the meetings'
MENTAL MODELS

that is so powerful that it cannot be completely encompassed by formal rules of inference. It follows, of course, that any theory that assumes that the logical properties of expressions derive directly from a mental logic cannot give an adequate account of those that call for a higher-order predicate calculus. This failure is a final and decisive blow to the doctrine of mental logic. The inference in the example above is readily made on the basis of a mental model in which the two predicates apply respectively to more than half the entities designated as musicians:

\[
\begin{align*}
m &= c \\
m &= c \\
m &= c = r \\
m &= r \\
m &= r
\end{align*}
\]

There is no way to satisfy the two premises that does not create an overlap.

One final example shows the importance of mental models in logical controversy. The philosopher Jaakko Hintikka (1974) has observed that the sentence:

Some relative of each villager and some relative of each townsmen hate each other

appears to be entirely synonymous with the corresponding sentence in which the noun phrases have been swapped round:

Some relative of each townsmen and some relative of each villager hate each other.

The order of the quantifiers in the corresponding expressions in the quantificational calculus does matter: there, the two expressions are not equivalent. Hintikka accordingly claimed that the sentences call for a logic with 'branching' quantifiers, i.e., quantifiers whose order is immaterial.

Both sentences are satisfied by the following model:

\[
\begin{align*}
v &\rightarrow r &\_\rightarrow t \\
v &\rightarrow r &\_\rightarrow t
\end{align*}
\]

where the \(v\)'s denote villagers, the \(t\)'s denote townsman, the \(r\)'s denote other people, the dotted line stands for the relation of being a relative of someone, and the solid line stands for the relation of mutual hatred. Hintikka's claim is debatable, but the moral I want to draw does not depend on whether he is right or wrong. It is that the only way to tease out the logic of such sentences is to construct models of them. The cases that are genuinely problematical to modern logicians naturally elicit the
same strategy that was employed by Aristotle. The logician first relies on semantic intuitions to make pre-theoretical judgments about inferences, and only subsequently sets up formal machinery to codify those judgments. If there are assertions that require branching quantifiers, there is no doubt that inferences that depend on them can be made by recourse to mental models. The question of the validity of such inferences depends initially on the exercise of intuitions, and this exercise, I contend, relies on the mobilization of mental models or their external counterparts.

The acquisition of explicit inferential ability

The mystery of how the ability to reason develops is so perplexing that some theorists, like Piaget, consider that children think in an essentially different way from adults, whereas others take the diachronically opposite point of view and postulate an innate logical ability. Perhaps the chief proponent of the latter doctrine is Jerry Fodor (1975, 1980). The essence of his argument is that there are and can be no theories of how concepts are acquired ab initio, but only theories that might explain how people can build up new beliefs inductively on the basis of the concepts they already possess. In particular, a child at a stage of development corresponding, say, to propositional logic, could not pass by way of learning to a stage corresponding to a more powerful logic such as the quantificational calculus. This transition is impossible, says Fodor.

Why? Because to learn quantificational logic we are going to have to learn the truth conditions of such expressions as $(x)Fx$ [i.e. for any x, x is F]. And, to learn those truth conditions, we are going to have to formulate, with the conceptual apparatus available at stage 1, some such hypotheses as $(x)Fx$ is true if and only if... But of course, such a hypothesis can't be formulated with the conceptual apparatus available at stage 1; that is precisely the respect in which propositional logic is weaker than quantificational logic...

If you think about this, you will see that this is an entirely general form of argument, one that shows that it is never possible to learn a richer logic on the basis of a weaker logic, if what you mean by learning is hypothesis formation and confirmation. Yet I say again that learning must be non-demonstrative inference [i.e. inductive inference]; there is nothing else for it to be. And the only model of non-demonstrative inference that has ever been proposed anywhere by anyone is hypothesis formation and confirmation. (Fodor, 1980, pp. 148–9)

The moral that Fodor draws is an extreme version of nativism – no concept is invented; all concepts are innate. Alas, any argument that purports to explain the origins of all intellectual abilities by postulating that they are innate merely replaces one problem by another. No one knows how deductive competence could have evolved according to the principles of neo-Darwinism.

In fact, the claim that there is no way in which the logical power of the mind could be increased by learning is ambiguous. Any computable function can be characterized in terms of the three sorts of primitive recursive functions (the zero, identity, and successor functions) and the three operations that create new functions out of old (composition, recursion, and minimization – see Chapter 1). Hence, any computable procedure whatsoever can be constructed out of these functions. One way in which to think of learning is as a procedure that discovers a way in which to combine old functions so as to create new ones. The system starts with an inborn set of functions and an inborn set of procedures for combining them, and learning consists in constructing functions that the system did not hitherto possess out of the functions that it did possess. There is accordingly both a sense in which the system does not increase in power – it always had the power of a Universal Turing machine – and another sense in which it does increase in power – hitherto it was unable, say, to make a transitive inference, but now it can do so. We can see why it is not necessary to accept Fodor’s conclusion. First, there is no need to assume that stage 1 is weaker in computational power than stage 2: in stage 1, an individual has a weaker logic than in stage 2, but does possess the capacity to construct a more powerful logic. Second, one must not confuse what is learned with how it is learned. When Pythagoras first proved his celebrated theorem, he learned something new – a certain proposition, which he may have believed to be true, was necessarily true. How Pythagoras discovered his proof may indeed have depended on all sorts of inductive procedures, but this supposition in no way infers what Pythagoras learnt. In general, it may be true, as Fodor claims, that learning must be a process of non-demonstrative inference, but it does not follow that all that can be learned is the inductive confirmation, or disconfirmation, of a belief. To argue otherwise is to argue that one can never learn necessary truths – they are all, as Plato supposed, inborn.

The real problem of learning is to explain the discovery of useful ways of combining old functions. Imagine an artificial learning system based on the standard recursive functions, which proceeds by using composition, recursion, and minimization, first on the set of primitive functions and then on the functions that it has created in this way. The system can in principle construct any of the denumerably infinite number of computable functions, but, like the volumes in Borges’s library of Babel, most of these
functions will be useless. The system has to be guided by a set of heuristics that determine what is useful. Given such heuristics, even a process of trial and error could learn how to do addition and multiplication by combining the basic recursive functions. (Such a process has indeed been implemented in a computer program by David Fallside, a former student of mine at Sussex.) The mind almost certainly has a rich set of native functions, together with specific procedures for guiding the process of constructing new functions out of old. Likewise, just as the environment may select among innately specified concepts (Fodor, 1980, p. 151), it may also help to place constraints on the combinations of functions that the mind pursues.

The conclusion to be drawn is not that nativism should be abandoned – there must be an innate armamentarium of data and procedures. And Fodor is right to argue that stages in inferential ability cannot possibly be associated solely with mental logics of increasing power. Where he is wrong, however, is in his stronger thesis that in principle all concepts are innate and that inferential skill cannot be learned.

Because the theory of mental models has no need of mental logic, it does not have to explain the initial development of reasoning ability by way of the acquisition of rules of inference. What children learn first are the truth conditions of expressions: they learn the contribution of connectives, quantifiers, and other such terms to these truth conditions. And, until they have acquired this knowledge about their language, they are in no position to make verbal inferences. Once they have learned such truth conditions, there may still be impediments that prevent them from realizing their full inferential competence. The reason that 12-year-olds are unable to cope with more complex syllogisms may depend, for example, on the limited processing capacity of their working memories. This system has yet to develop to its greatest power.

**Conclusions**

In this chapter, I have outlined a general theory of inference based on mental models, and I have established the following points:

1. The theory embraces both implicit and explicit inferences. Implicit inferences depend on constructing a single mental model; explicit inferences depend on searching for alternative models that may falsify putative conclusions. Since the interpretation of premises as mental models provides a powerful representational system, the theory is also extensible to most, if not all, classes of arguments deployed by speakers of natural language.
2. The theory solves the central paradox of how children learn to reason, namely, how they could acquire rules of inference before they were able to reason validly. The paradox is resolved because it rests on a false assumption: children need neither acquire rules of inference (pace Piaget) nor possess them innately (pace Fodor) in order to make valid deductions. It is possible to reason validly without logic.
3. The theory is entirely compatible with the fact that human beings are capable of making valid deductions.
4. It is also compatible with the origins of logic. It assumes that people make inferences without recourse to a mental logic. Certain inferences cause them difficulty, however, and therefore provide the motivation for the search for systematic principles governing validity.

These four points correspond to the criteria that any explanatory adequate theory of reasoning must satisfy. After I had introduced them in Chapter 4, I pointed out that the fundamental dilemma yet to be resolved by any theory of reasoning was to allow for rationality in ideal conditions and for human error in less than ideal conditions. The theory of mental models dispels this dilemma. The remaining lacunae in the theory concern how sentences containing complex quantifiers or several quantifiers are translated into mental models; how these models are manipulated in searching for counter-examples; and what constraints apply to the set of possible models – cf. the notation of mental models, which was quietly expanded in order to deal with the membership of one class within another.

The first of these gaps calls for an effective procedure. There are many feasible algorithms for translating complex and multiple quantifiers into models, and for manipulating those models; and a search for models that are counter-examples to putative conclusions will either succeed or fail in a finite number of steps granted that any mental model is finite in size. (Reasoning about infinite sets such as the natural numbers calls for a different approach that will be described later in the book.) Hence there is no doubt that appropriate algorithms exist. The problem of determining which algorithm is used in inference is difficult to solve because people who are untrained in logic find these particular inferences very hard.

Before the second of these gaps can be filled, and the set of possible mental models delineated (in Chapter 15), it will be necessary to examine both the way in which models differ from other representations and the process of understanding natural language.