Précis of Deduction

Philip N. Johnson-Laird
Department of Psychology, Princeton University, Princeton, NJ 08544
Electronic mail: phil@clarity.princeton.edu

Ruth M. J. Byrne
Department of Psychology, Trinity College, University of Dublin, Dublin 2, Ireland
Electronic mail: rbyrne@wexf1.tcd.ie

Abstract. How do people make deductions? The orthodox view in psychology is that they use formal rules of inference like those of a 'natural deduction' system. Deduction argues that their logical competence depends, not on formal rules, but on mental models. They construct models of the situation described by the premises, using their linguistic knowledge and their general knowledge. They try to formulate a conclusion based on these models that maintains semantic information, that expresses it parsimoniously, and that makes explicit something not directly stated by any premise. They then test the validity of the conclusion by searching for alternative models that might refute the conclusion. The theory also solves long-standing puzzles about reasoning, including how syllogistic reasoning occurs in daily life. The book reports experiments on all the main domains of deduction, including inferences based on propositional connectives such as 'if' and 'or,' inferences based on relations such as 'in the same place as,' inferences based on quantifiers such as 'none,' 'any,' and 'only,' and metalinguistic inferences based on assertions about the true and false. Where the two theories make opposite predictions, the results confirm the model theory and rule out the formal rule theories. Without exception, all of the experiments corroborate the two main predictions of the model theory: inferences requiring only one model are easier than those requiring multiple models, and erroneous conclusions are usually the result of constructing only one of the possible models of the premises.

Keywords: conditionals; deduction; formal rules; mental models; syllogistic reasoning; quantifiers; rationality; reasoning; Allegro; theorem proving

I'm thirsty, he said. I have sevenpence. Therefore 1 buy a pint.

The conclusion of your syllogism, I said lightly, is fallacious, being based on licensed premises.

Plato, O'Brien. A Saum-Two-Birds. (1899, p. 29)

From long habit the train of thoughts ran so swiftly through my mind that I arrived at the conclusion without being conscious of intermediate steps. There were such steps, however. The train of reasoning ran, "There is a gentleman of a particular type, but with the air of a military man. Clearly, an army doctor, then. He has just come from the tropics, for his face is dark, and that is not the natural tint of his skin for his wrists are fair. He has undergone hardship and sickness, as his haggard face says clearly. His left arm has been injured. He holds it in a stiff and unnatural manner. Where is the tropics could an English army doctor have seen much hardship and got his arm wounded? Clearly in Afghanistan. The whole train of thought did not occupy a second. I then remarked to you came from Afghanistan, and you were astonished.


1. Introduction

A complete theory of thinking has to explain calculation, deduction, induction, creation, and the association of ideas. In this book we set ourselves a more modest goal: to explain the nature of deduction and to characterize its underlying mental processes. Why deduction? One reason is its intrinsic importance: it plays a crucial role in many tasks. A world without deduction would be a world without science, technology, laws, social conventions, and culture. And if you want to dispute this claim, we shall need to assess the validity of your arguments. Another reason for studying deduction is that after eighty years of psychological experiments on the topic it ought to be ripe for solution. In the introductory chapter, we provide a brief but necessary background in logic before we plunge into the murky problems of psychology. We describe a semantic method for deduction in the propositional calculus and explain why no practical procedures can examine models in the predicate calculus. Logicians such as Beth (1955/1960), Hintikka (1956), and Smullyan (1968) have proposed formal systems based on the idea of a search for counterexamples.

2. The cognitive science of deduction

The degree to which people are logically competent is a matter of dispute. One view is that they never make a logical error: deduction depends on a set of universal principles applying to any content, and everyone exercises these principles infallibly. They merely forget the premises sometimes, or make unwarranted assumptions. Infallibility seems so contrary to common sense that, as you might suspect, it has been advocated by some philos-
It has also been advocated by some psychologists (e.g., Heine 1978). Others take a much darker view about logical competence and proceed theories that render logical theories intrinsically irrational (e.g., Ercleison, 1974). Our view is that people are rational, in principle, but fallible in practice. Even though they are not formally taught logic, they develop the ability to make valid deductions, that is, to draw conclusions that must be true given that the premises on which they are based are true. Moreover, they sometimes know that they have made a valid deduction. They also make deductive errors in certain circumstances and are even prepared to concede that they have done so (Wason & Johnson-Laird 1972). These metatemporal intuitions are important because they provide the way for the invention of self-conscious methods for checking validity, that is, logic. Yet logic would hardly have been invented if there were never occasions when people were uncertain about the status of an inference.

When people reason deductively, they start with some information — either evidence of the senses or a verbal description — and they assess whether a given conclusion follows logically from this information. In real life there is often no given conclusion, so they generate a conclusion for themselves. Logic alone is insufficient to characterize intelligent reasoning in this case, because any set of premises yields an infinite number of valid conclusions. Most of them are baseless, such as the conjunction of a premise with itself, and so some individuals, apart from a logician, would dream of drawing such conclusions. Hence, when individuals make a deduction in daily life, they must be guided by more than logic. They draw useful conclusions. The evidence suggests that they tend to move some information to the premises, to reexpress it more parsimoniously, and to establish something not directly asserted in a premise. If nothing meets these constraints, they declare that there is no valid conclusion.

What are the mental mechanisms underlying deduction? Cognitive scientists have put forward theories based on three distinct ideas:

1. formal rules of inference,
2. content-specific rules of inference, and
3. mental models.

A number of concepts, of course, can pick out one theory against all others, because infinitely many theories are compatible with any finite set of observations. Our problem is simpler: it is to decide amongst existing theories of these three sorts. One of them, however, is not a fully independent option. A content-specific rule such as:

\[ \text{If } p, \text{ then } q \]

or a pragmatic reasoning schema (Cheng & Holyoak 1985) such as:

\[ \text{If the action is to be taken, then the precondition must be satisfied.} \]


can only be part of a general inferential system. Like their logical cousins, meaning postulates, these content-specific rules require additional inferential machinery if the theory is to account for deductions that do not depend on factual knowledge. Hence, the general theoretical possibilities reduce to two formal rules or mental models. Philosophers, psychologists, linguists, and artificial intelligence investigators have long assumed that the mind contains formal inference rules. They have characterized these rules in ways akin to the logical method of “natural deduction” (see, e.g., Braine 1978; Johnson-Laird 1973; Macnamara 1966; Osherson 1973; Pollock 1989; Reite 1973; Bips 1983; Sperber & Wilson 1986). Each connective, such as “or,” “and,” and “or,” has its own rule. Deduction accordingly consists in representing the logical form of premises and then using the formal rules of inference to try to find a derivation of the conclusion from the premises. If no derivation of the conclusion can be found, reasoners will respond that the inference is invalid.

Here is an example of how a formal rule theory works. When people reason from conditionals, they are readily able to make a modus ponens deduction:

\[ \text{If there is a triangle then there is a circle.} \]
\[ \text{There is a triangle.} \]
\[ \text{Therefore, there is a circle.} \]

but they are less able to make the modus tollens deduction:

\[ \text{If there is a triangle then there is a circle.} \]
\[ \text{There is not a circle.} \]
\[ \text{Therefore, there is not a triangle.} \]

Indeed, many intelligent individuals say that nothing follows in this case. Theorists postulate that modus tollens is easy because there is a corresponding formal rule in mental logic:

\[ \text{If } p \text{ then } q \]
\[ \text{not- } q \]
\[ \therefore \text{ not- } p \]

Modus tollens is harder because there is no rule for it, and so it calls for a derivation:

\[ \text{If } p \text{ then } q \]
\[ \text{not- } q \]
\[ \therefore \text{ not- } p \]

(by hypothesis)

\[ \therefore q \]

(by modus ponens)

\[ \therefore q \text{ and not- } q \]

(by conjunction)

\[ \therefore \text{ not- } p \]

(by reductio ad absurdum)

In general, formal rule theorists predict that the difficulty of a deduction depends on two factors: the length of the formal derivation and the availability or ease of use of the relevant rules (see, e.g., Braine et al. 1984; Bips 1983).

In contrast, valid deductions can be made in the propositional calculus by manipulating truth tables, but logically untrained individuals are unlikely to use this method because it calls for too much information to be kept in mind. To abandon truth tables, however, is not necessarily to abandon a semantic approach to reasoning. The mental-model theory assumes that people reason from their understanding of a situation and that their starting point is accordingly a set of models—typically a single model for a single situation—that is constructed from perceiving the world or from understanding discourse, or both (Johnson-Laird 1983). Mental models may occur as visual images, or they may not be accessible to consciousness. What matters are their structures, which are identical to the structures of the states of affairs, whether perceived or conceived, that the models represent. Models also make as little as possible explicit be-
Johnson-Laird & Byrne: Deduction

In contrast, the modus tollens premise:

There is not a triangle

eliminates the explicit model to leave only the implicit model, from which nothing seems to follow. The deduction can be made only by fleshing out the models of the conditional, for example:

\[
\begin{align*}
\{&0\} \quad \{&\Delta\} \\
\{&\neg 0\} \quad \{&\Delta\} \\
\{&\neg \neg \Delta\} \quad \{&\neg \Delta\}
\end{align*}
\]

where "\(\neg\)" represents negation. The premise, "There is not a triangle," now eliminates the first two models to yield the conclusion:

There is not a circle.

The difference in difficulty between the two deductions, according to the rule theories, arises from the lengths of their derivations. This hypothesis fails to explain why the difference disappears when the conditional premise is expressed using "only if" (Evans 1977):

There is a circle only if there is a triangle.

If people used the rule for modus ponens, then the difference in difficulty should swap round – granted, as formal theorists assume (Braine 1978), that the premise is equivalent to:

If there isn't a triangle then there isn't a circle.

In contrast, the model theory postulates that the "only if" premise leads to the construction of explicit models of both the affirmative antecedent and the negated consequent:

\[
\begin{align*}
\{&0\} \quad \{&\Delta\} \\
\{&\neg 0\} \quad \{&\neg \Delta\}
\end{align*}
\]

and so both deductions are of the same difficulty. The rule theory can be altered post hoc to accommodate this phenomenon, but there are a number of other results that presently defy explanation in terms of rules, for example, the greater ease of deductions based on exclusive disjunction (two models) than those based on inclusive disjunction (three models).

The model theory has been implemented in a computer program, and it has led to novel predictions of its own. It correctly anticipated, for example, that modus tollens would be easier with a biconditional (two models) than with a conditional (two or three models). It also predicted the striking difficulty of "double disjunctions" and the sorts of error that occur with these problems (as we showed in a series of experiments carried out in collaboration with Walter Schaeken of the University of Leuven; see Johnson-Laird et al. 1992). Double disjunctive premises such as:

Linda is in Cannes or Mary is in Tripoli, or both.
Mary is in Havana or Cathy is in Sofia, or both.

call for five models:

\[
\begin{align*}
\{&c\} \quad \{&t\} \quad \{&s\} \\
\{&c\} \quad \{&b\} \quad \{&s\} \\
\{&c\} \quad \{&b\} \\
\{&c\} \quad \{&t\} \quad \{&s\}
\end{align*}
\]
where "a" denotes Linda in Cannes, "b" denotes Mary in Tripoli, "h" denotes Mary in Havana, and "s" denotes Cathy in Sofia. A typical conclusion is:

Linda is in Cannes and Cathy is in Sofia and Mary may be in Tripoli.

It is based on only some of the possible models of the premises, [e] [f] [g] and [e] [s], and it is invalid because other models falsify it, for example, [e] [h]. Rule theories, however, have yet to lead to the discovery of novel phenomena; adherents have fitted their theories to data from variated sets of deductions, typically using one parameter for each rule of inference (Braine et al. 1984; Rips 1983). A reexamination of these results shows that the model theory provides an equally plausible account of them, and in some cases goes beyond rule theories in its explanatory power.

4. Conditionals

Although attempts have been made to develop rule theories for connectives that do not occur in formal logic (Rips 1983), a major problem for these accounts is the lack of uniform logical properties for many connectives. Similarly, some indicative conditions are truth-functional, that is, they have meanings equivalent to a truth-table definition, whereas others appear not to be. Some are interpreted as biconditionals and some as conditionals (Legrenzi 1970). The model theory accommodates all of them. Those with "defective" truth tables have an implicit model of the state in which the antecedent is false; those that are truly truth-functional have explicit models of the state in which the antecedent is false. Hence, conditionals have a simple semantics based on mental models.

Counterfactual conditionals, such as:

If pigs had no teeth, they would gum you to death cannot be truth-functional because antecedent and consequent are both false. Theories based on formal rules therefore have little to say about them, but we show how their meanings can be mentally represented by models of actual and counterfactual states, and how a semantic theory of causal relations (Miller & Johnson-Laird 1970) dovetails with this account.

Models can be interrelated by a common reference or by general knowledge. Byrne (1989) demonstrated that these relations in turn can block modus ponens. As the model theory predicted, when subjects were given a pair of conditionals of the following sort:

If Lisa goes fishing, then Lisa has a fish supper.
If Lisa catches some fish, then Lisa has a fish supper.

and the categorical assertion:

Lisa goes fishing.

they tended not to conclude:

Lisa has a fish supper.

The second conditional reminded them that Lisa also needs to catch some fish. The suppression of the deduction shows that people do not have a secure intuition that modus ponens applies equally to any content. Yet, this intuition is a criterion for the existence of formal rules in the mind. The model theory also predicted the sorts of sentences that are likely to be paraphrased by conditionals. They are, as we confirmed experimentally, those that describe an outcome as a possibility, because a possibility tallies with the implicit model in the set for a conditional.

A major problem for formal rule theories is that reasoning is affected by the content of deductive problems. The best-known illustration is provided by Peter Wason's selection task (Wason 1966; 1952; Wason & Johnson-Laird 1972). In the original version of the task, four cards are put in front of a subject, bearing on their uppermost face a single symbol: A, B, 2, and 3; and the subjects know that every card has a letter on one side and a number of the other side. Their task is to select just those cards they need to turn over in order to determine whether the following conditional rule is true or false:

If a card has an A on one side then it has a 2 on the other side.

The majority of subjects select the A card, or the A and the 2 cards. Surprisingly, they fail to select the card corresponding to the case where the consequent is false: the 2 card. Yet, the combination of an A with a 2 falsifies the rule.

The selection task has generated a large literature, which is not easy to integrate, and one investigator, Evans (1989), has even wondered whether it tells us anything about deduction as opposed to heuristic biases. He argues that subjects make those selections that merely match the cards mentioned in the rule. Hence, when the rule is negative:

If there is an A then there is not a 2.

many subjects correctly select the 2 card (which falsifies the consequent). However, realistic conditional rules, such as:

If a person is drinking beer then the person must be over 18.

have a striking effect on performance. The subjects tend to make the correct selections of the cards corresponding to the true antecedent and the false consequent (e.g., Cheng & Holyoak 1985; Griggs 1983; Griggs & Cox 1982).

Theories based on formal rules, as Manktelow and Over (1987) have argued, cannot easily account either for the failure to select the false consequent in the original task or for its selection with realistic conditionals. There is no difference in the logical form of the two sorts of conditionals that could account for the results. Moreover, these arch-formalists, the Piagetians, claim that children have a capacity for falsification as soon as they attain the level of formal operations (Inhelder & Piaget 1958). Piaget describes this ability in the following terms: to check the truth of a conditional, if p then q, a child will look to see whether or not there is a counterexample, p and not-q (see Beth & Piaget 1966, p. 181). Yet adults conspicuously fail to do so in the original version of the selection task.

Several reasons have been put forward to explain why a realistic conditional may elicit the correct selection. They are all variants on the theory that people use context-specific rules of inference. Thus, Cheng and Holyoak (1985) have proposed "pragmatic reasoning schemas", which are rules of inference induced from experience with causation, permission, and obligation. The permission schema, for example, contains four rules: (1) If the action is to be taken, then the precondition must be
satisfied. (2) If the action is not to be taken, then the precondition need not be satisfied. (3) If the precondition is satisfied, then the action may be taken. (4) If the precondition is not satisfied, then the action must not be taken. The conditional about beer drinking cues the schema and the fourth rule leads to the selection of the card corresponding to the false consequence. Conditionals about arbitrary letters and numbers cannot normally elicit such schemas.

Other experimental manipulations lead to insight into the selection task even though they do not depend on general knowledge, for example, the use of simpler rules, such as "All the triangles are white" (Wason & Green 1960). They illustrate how the use of simpler rules can lead to insight into the task. The model theory explains the selection task in a different way: (1) People reason only about what is explicitly represented in their models - in this case, their models of the rule. (2) They select from the explicitly represented cards those for which the hidden value could have a bearing on the truth or falsity of the rule, that is, those that are represented exhaustively in their models of the rule. Hence, any manipulation that leads to the falsifying of the model of the conditional with explicit representations of the false consequence will tend to yield insight into the task.

The conditional, "If there is an A on one side, then there is a 2 on the other side," yields a model containing only the cards mentioned in the rule:

\[
\begin{align*}
\text{[A]} & \quad \text{2} \\
\end{align*}
\]

or:

\[
\begin{align*}
\text{[A]} & \quad \text{[2]} \\
\end{align*}
\]

and so people tend to select only the A card, or the A and the 2 card. The model theory is thus compatible with Evans's matching bias on the assumption that negation leads to falsifying their assumptions.

The model theory is thus compatible with Evans's matching bias on the assumption that negation leads to falsifying their assumptions.

If a man has a tattoo on his face, then he eats cassava root tends to elicit selections of the following cards: no tattoo, and eats cassava root. There is a simple alternative explanation for this result. The subjects treat the rule as

\[
\begin{align*}
\text{A man may eat cassava root only if he has a tattoo.}
\end{align*}
\]

Such an assertion, as we argued earlier, calls for models of the following sort:

\[
\begin{align*}
\text{[eats cassava]} & \quad \text{tattoo} \\
\end{align*}
\]

\[
\begin{align*}
\text{→ eats cassava} & \quad \text{[→ tattoo]} \\
\end{align*}
\]

The cards which bear on the truth or falsity of the rule are accordingly: no tattoo, and eats cassava, and so the subjects will tend to select them. There is no need to postulate a specific inferential module concerning the violation of social contracts.

5. Reasoning about relations

In a rule theory, the logical properties of a relation have to be stated in postulates or content-specific rules. "In the same place as," for instance, is a transitive relation, and this logical property can be captured in the postulate:

\[
\begin{align*}
\text{For any } x, y, z, \text{if } x \text{ is in the same place as } y \text{ and } y \text{ is in the same place as } z, \text{ then } x \text{ is in the same place as } z.
\end{align*}
\]

The model theory does not need such postulates. Their work is done by a representation of the meaning of the relation, that is, its contribution to truth conditions. One advantage is that the logical properties of relations emerge from their meanings. It is then easy to see why certain relations have properties that are affected by the mental model of the situation under discussion. Thus, "to the right of" calls for an indefinite number of different degrees of transitivity. The premises:

\[
\begin{align*}
\text{Matthew is to the right of Mark.} \\
\text{Mark is to the right of John.}
\end{align*}
\]

lead naturally to a transitive conclusion:

\[
\begin{align*}
\text{Matthew is to the right of John}
\end{align*}
\]

provided that the seating arrangement resembles Leonardo Da Vinci's painting of The Last Supper. The conclusion may be blocked, however, if the individuals are seated at a round table. The degree of transitivity then depends on the radius of the table and the proximity of the seats. The logical properties of the relation require an indefinite number of different meaning postulates - one for each degree of transitivity. Yet if the meaning of such expressions is specified as a direction from a reference individual:

\[
\begin{align*}
\text{to the front} & \quad \uparrow \\
\text{to the left} & \quad \odot \\
\text{to the right} & \quad \downarrow \\
\text{to the back} & \quad \nabla
\end{align*}
\]

then the degree of transitivity is an automatic consequence of the seating arrangement.

Simple transitive deductions led to an irresolvable controversy about the underlying inferential mechanisms (cf. Clark 1969; Huttonlocher 1968). However, a computer program implementing the model theory of spatial reasoning revealed an unexpected difference between models and rules. Certain spatial deductions require just one model yet call for complex derivations based on rules; other deductions require multiple models yet call for simple derivations based on rules. The difference arises because objects interrelated in a single model may not have occurred in the same premise and so a formal procedure needs to derive the relation between them. In contrast, objects may be interrelated in a single premise.
and so a rule can be immediately applied to them and yet the description as a whole may be compatible with more than one possible model. Consider, for example, the following problem:

The triangle is on the right of the circle.  
The cross is on the left of the circle.  
The asterisk is in front of the cross.  
The line is in front of the triangle.  
What is the relation between the asterisk and the line?

The description corresponds to a single determinate model:

* □ △

Hence, it should be relatively easy to answer:

The asterisk is on the left of the line.

If one word is changed in the second premise, the result is the following problem:

The triangle is on the right of the circle.  
The cross is on the left of the triangle.  
The asterisk is in front of the cross.  
The line is in front of the triangle.  
What is the relation between the asterisk and the line?

This description yields at least two distinct models:

+ ⋇ △

but both models yield the same conclusion:

The asterisk is on the left of the line.

The model theory predicts that the task should be harder because it calls for multiple models.

Rule theories for spatial reasoning, such as the one proposed by Hagert (1984), need rules for transitivity and rules for two-dimensional relations, such as:

Left (x, y) & Front (z, x) → Left (front (z, x), y)

where the right-hand side of the rule signifies “z is in front of x, which is on the left of y.” Whatever the form of the rules, the one-model problem requires a derivation of the relation between the cross and the triangle, whereas this relation is directly asserted by the second premise of the multiple-model problem. Hence, the rule theory predicts that the one-model problem should be harder than the multiple-model problem, which is exactly the opposite to the prediction made by the model theory. We carried out a series of experiments on spatial reasoning (Byrne & Johnson-Laird 1989), and the results corroborated the model theory but ran counter to rule theories: inferences were harder if they called for the construction of multiple models.

6. Syllogistic reasoning

The most powerful forms of deduction depend on quantifiers, such as “all,” “some,” and “none.” When assertions contain only a single quantified predicate, they can form the premises of syllogisms, such as:

All the athletes are bodybuilders.  
All the bodybuilders are competitors.  
∴ All the athletes are competitors.

A syllogism has two premises and a conclusion in one of four “moods” shown here:

All A are B  (a universal affirmative premise)  
Some A are B  (a particular affirmative premise)  
No A are B  (a universal negative premise)  
Some A are not B  (a particular negative premise)

To support a valid conclusion the two premises must share a common term (the so-called middle term), and hence the premises can have four different arrangements (or “figures”) of their terms:

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

The syllogism above is in the first of these figures (where A = athletes, B = bodybuilders, and C = competitors).

Given that such premises can be in one of four moods, there is a total of 64 distinct forms of premises. Scholastic logicians recognized that the order of the premises had no logical effect so they adopted the convention that the subject of the conclusion was whichever end term occurred in the second premise. At first, psychologists followed Scholastic logic; as a result, they ignored half of the possible forms of syllogism. The early studies were also vitiated by methodological flaws. Subjects could use guessing and other noninferential processes, because they had only to evaluate given conclusions. In the 1970s, however, we asked subjects to generate their own conclusions, and this procedure enabled all 64 forms of premises to be investigated (Johnson-Laird & Huttonlocher, reported in Johnson-Laird 1975). One result was the discovery of a very robust “figural” effect.

In general, a syllogism in the figure:

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

tends to elicit conclusions of the form:

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

whereas a syllogism in the figure:

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

tends to elicit conclusions of the form:

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

This bias is probably a result of the order in which information is combined in working memory: conclusions are formulated in the same order in which the information is used to construct a model. Alternatively, the bias may reflect a pragmatic preference for making the subject of a premise into the subject of the conclusion (Wetherick & Gilhooly 1990). This linguistic bias, however, fails to explain the progressive slowing of responses over the four figures shown above, or the increasing proportion of “no valid conclusion” responses. The phenomena are accounted for by the working memory hypothesis, according to which there is both a recoring of information in a premise and a recoring of the premises themselves to bring two occurrences of the middle term into temporal contiguity (Johnson-Laird & Bana 1984).

No one has proposed a complete psychological theory of syllogistic inference based on formal rules, perhaps because the lengths of formal derivations for valid syllogisms fail to account for differences in difficulty amongst...
them. One long-standing proposal, however, is that reasoners tend to match their conclusions to the mood of one or other of the premises (Begg & Denny 1983; Woodward & Sells 1985). This notion of an "atmosphere" effect continues to exert its influence on recent theories (e.g., Madsen 1994; Polk & Newell 1988), but we have observed a phenomenon that is damaging to all versions of the atmosphere hypothesis (see Johnson-Laird & Byrne 1989). When both premises of a syllogism were based on the quantifier "only," just 16% of conclusions contained it, whereas 45% of conclusions contained "all." Likewise, where one premise was based on "only," just 2% of conclusions contained it. In our view, the apparent evidence supporting the atmosphere hypothesis derives, in fact, from the natural consequences of building models based on the meaning of the premises and then using a procedure to construct parsimonious conclusions. The bias towards "all" corroborates our assumption that "only" elicits explicit negative information.

The main controversy about syllogisms is about the nature of models that represent the premises: are they Euler circles (Erickson 1974), Venn diagrams (Newell 1951), or some other format (Guyote & Sternberg 1981)? We argue that models represent finite sets of entities by finite sets of mental tokens rather than by circles inscribed in Euclidean space. This hypothesis correctly predicts two of the most robust results in syllogistic reasoning. First, syllogisms that call for only one model of the premises are reliably easier than those that call for multiple models. We have yet to test an individual who does not conform to this prediction. Second, erroneous conclusions tend to correspond to descriptions of a subset of the models of the premises—typically just one of the models (as in the case of propositional reasoning). We have also corroborated this finding in a study of subjects' memory for conclusions they had drawn. Even when they had correctly responded that there was no valid conclusion, if they later thought they had drawn one, it was invariably the one the theory predicts they had initially constructed, only to reject because it was refuted by an alternative model (Byrne & Johnson-Laird 1990).

A reasoner's goal is to reach true, or at least plausible, conclusions rather than merely valid ones. Knowledge can assist this process by providing pertinent information and a means for assessing the truth of conclusions. You are likely to judge that a conclusion is true if it corresponds to the state of affairs in the world, or if it coheres with your other beliefs. Can knowledge directly affect the process of reasoning? The issue is highly controversial (see Nisbett & Ross 1980). If reasoning is based on formal rules, it cannot be affected by beliefs: formal rules are, by definition, blind to the content of premises. But the theory of mental models predicts such effects: individuals who reach a putative conclusion that fits their beliefs will tend to stop searching for alternative models that might refute their conclusion.

We examined this prediction experimentally in collaboration with Jane Oakhill and Alan Garnham of the University of Sussex (Oakhill & Johnson-Laird 1985b, Oakhill et al. 1989). When we gave intelligent but logically untrained individuals the following premises, for example:

All of the Frenchmen in the room are wine-drinkers.
Some of the wine-drinkers in the room are gourmets.

the majority of them drew the conclusion:
Some of the Frenchmen are gourmets.
But, when we gave the subjects the premises:
All of the Frenchmen in the room are wine-drinkers.
Some of the wine-drinkers in the room are Italians.

hardly any of them drew the conclusion:
Some of the Frenchmen in the room are Italians.

and most people responded correctly that there is no valid conclusion (interpreting the end terms). This phenomenon confirms that knowledge influences the process of deduction. Reasoners evidently construct an initial model that supports a putative conclusion. If the conclusion fits their beliefs, the process of inference halts; if it does not fit their beliefs, the process of inference continues to search for an alternative model that refutes it.

Images are a special case of models, but models can also contain conceptual tags to represent various sorts of abstract information that cannot be visualized. The best example is negation. The use of such annotations could perhaps be avoided by maintaining a linguistic representation of the premises, but our experiments provide evidence that reasoners represent negation directly in their models. The assertion:

All the athletes are bankers.

is represented by a model of the following sort:

\[ a \quad b \]
\[ a \quad b \]
\[ \neg a \quad b \]
\[ \neg a \quad \neg b \]

where "a" denotes an athlete, "b" denotes a banker, and each line in this diagram represents a different individual in the same model (unlike the propositional models that we presented earlier). The number of individuals remains arbitrary, but it is likely to be small. The brackets indicate that the a's have been exhaustively represented in relation to the b's. Hence, in fleshing out the implicit individual(s) represented by the three dots, if a's occur they must be accompanied by b's. One way in which to flesh out the model is as follows:

\[ a \quad b \]
\[ a \quad b \]
\[ \neg a \quad b \]
\[ \neg a \quad \neg b \]

where "\neg" represents negation.

The assertion:

Only the bankers are athletes.

has the same truth conditions as the assertion containing "all," but it makes explicit right from the start that anyone who is not a banker is also not an athlete. Hence, its initial model according to the theory is of the following sort:

\[ b \quad a \]
\[ b \quad a \]
\[ \neg b \quad \neg a \]

The implicit individual can be fleshed out as:

\[ b \quad \neg a \]

The models of the two assertions are therefore equivalent.
in content, but the equivalence is not immediately apparent to subjects because the initial model for "all" makes explicit just the affirmative information, whereas the initial model for "only" makes explicit both affirmative and negative information.

The model theory predicts that deductions based on what is explicit in a model should be easier than those that depend on fleshing out implicit information. It follows that the premises:

All athletes are bankers.
Mark is an athlete.
should readily yield the conclusion:
Mark is a banker.

whereas the premises:

All athletes are bankers.
Mark is not a banker.
should less readily yield the conclusion:
Mark is not an athlete.

In this case, the model has to be fleshed out with negative information about the set of individuals who are not bankers before the conclusion can be derived. The corresponding problems based on "only" yield a different prediction. There should be no difference between the premises:

Only bankers are athletes.
Mark is an athlete.

and:

Only bankers are athletes.
Mark is not a banker.

because the models contain explicit negative information right from the start. The results from our experiment corroborate the theory (Johnson-Laird & Byrne 1999).

Our experiments with "only" highlight an important feature of the model theory. The representation of premises depends on their meaning; the inferential procedure of searching for a counterexample is entirely general and can be applied to any sort of model. It follows that the theory can easily be extended to accommodate assertions that contain a new quantifier or connective. It is necessary only to describe the contribution of the new term to models of assertions containing it. Once this semantics has been specified, the reasoning procedure can operate on the models and there is no need for new rules of inference. The parsimony of the model theory contrasts with rule theories, which must describe both the meaning of the new term (its contribution to truth conditions) and its rules of inference.

7. Reasoning with multiple quantifiers

Here is a simple but robust result. When we presented subjects with the premises:

None of the circles is in the same place as any of the triangles.
All of the triangles are in the same place as all of the crosses.
the majority of these drew the valid conclusion:
None of the circles is in the same place as any of the crosses.

But, when we presented them with the premises:

None of the circles is in the same place as any of the triangles.
All of the triangles are in the same place as some of the crosses.
only a few drew the valid conclusion:
None of the circles is in the same place as some of the crosses.

Some of the crosses are not in the same place as any of the circles.

Why is there this difference in accuracy? No one has proposed a theory based on formal rules that accounts for the difference; indeed if such a theory is based on formal rules akin to those postulated for propositional reasoning, then, as we show, the two problems have derivations of exactly the same length. The premises cannot be represented by Euler circles or Venn diagrams, which can cope only with singly quantified assertions, yet they can be represented by a model, because the model theory readily generalizes to the representation of multiply quantified assertions.

According to the theory, the premises of the first problem yield a single model:

\[[O][O][C][[A][A][A]+[+][+]]\]

where the vertical barriers demarcate separate places and the three sets are each exhaustively represented by an arbitrary number of tokens. This model supports the conclusion:

None of the circles is in the same place as any of the crosses.

There are no alternative models of the premises that refute the conclusion, and so it is valid. The premises of the second problem support a similar initial model:

\[[O][O][C][[A][A][A]+[+][+]]\]

where the crosses are not exhaustively represented. This model supports the same conclusion as before, but now the search for an alternative model that refutes the conclusion will be successful:

\[[O][O][C]+[[A][A][A]+[+][+]]\]

The two models support the conclusion:

None of the circles is in the same place as some of the crosses.

or equivalently:

Some of the crosses are not in the same place as any of the circles.

The first problem calls for only one model, whereas the second problem requires multiple models and that is why there is a difference in difficulty between them.

We report a series of experiments on multiply quantified inference carried out in collaboration with Patrizia Tabbosi of the University of Bologna, Italy. Once again, the results confirmed the predictions of the model theory and rule out other explanations in terms of scope of quantifiers, matching strategies, or particular difficulties of one quantifier as opposed to another (Johnson-Laird et al. 1999). One-model problems were reliably easier than multiple-model problems; and, once again, the subjects' erroneous conclusions typically corresponded to only one model of multiple-model premises.
8. Metadeduction

Reasoners can know that they have made a valid deduction without this higher-level ability human beings could not invent logic, make deductions about other people's deductions, or devise psychological theories of reasoning. We examine what little is known about such abilities, distinguishing between metareasoning, which depends on an explicit reference to validity or truth and falsity, and metacognitive reasoning, which depends on reference to what oneself or others may be deducing.

Rips (1989) has pioneered the investigation of metacognitive deduction using "knight and knave" puzzles, for example:

There are two sorts of people:

Knights always tell the truth; knaves always lie.
A asserts that C is a knave.
B asserts that C is a knave.
C asserts that A is a knight and B is a knave.
What are A, B, and C?

Rips develops a formal rule theory that offers an explanation of how subjects solve such problems. Our first concern is that such formal theories do not reflect the importance of truth and falsity: without them, there can be no notion of validity and no way to consider the completeness, in the logical sense, of formal rules. We accordingly accepted Rips's challenge to show how a mental model theory could also account for performance (Johnson-Laird & Byrne 1980). Rips also assumes that subjects adopt the same strategy for all such problems, which is based on deriving contradictions from hypotheses. We argue instead that subjects are likely to develop different strategies depending on the particular problems they encounter. We implemented a variety of these strategies in a computer program that reasons with models. One such strategy, for example, assumes that reasoners notice in the problem above that A and B both make the same assertion, and so they are either both knights or else both knaves. C, however, does not make the same assertion about both of them, and so C is a knave. Both A and B say so, and so they are both knights. Our theory accounts for certain experimental results that the formal theory leaves unexplained.

Psychology is a "recursive" discipline because a plausible theory of high-level cognition should reveal how the theory itself could have been created as a result of the theorist's high-level cognition. A theory of metareasoning should therefore provide some insight into its own development. Our theory postulates a capacity to think about thinking — to reflect on patterns of deduction and the preservation of truth, on what one has deducted for oneself, and on the implications of what others can deduce. This general metacognitive capacity enables people to construct models of thought, and to construct models of those models, and so on, recursively. In this way, simple reasoning strategies can be invented by logically untutored individuals. The same ability can be used by logicians to create formal calculi for deduction, and then to reflect upon the relations between these calculi and their semantics. And, most important, the ability can be used by cognitive scientists to construct theories about itself.

9. Models in computer programs for reasoning

A program for reasoning on the basis of models calls for three principal components. First, it must be able to interpret premises expressed in a subset of natural language, and to construct an appropriate set of models for them. Second, the program must be able to use these models to formulate a conclusion — a parsimonious conclusion that makes explicit information that was not expressed in any single premise. Third, the program must be able to search for alternative models of the premises in order to test validity.

We show how we have implemented all these components in programs. The first stage in constructing a model of a premise consists in a compositional interpretation of its meaning (i.e., intension). This calls for a grammar and a lexicon that both contain grammatical and semantic information and a parser that uses this information to combine the meanings of constituents according to the grammatical relations amongst them. The significance of a premise — the particular proposition it expresses — depends on a number of additional factors, particularly on its context of use. Context in everyday discourse is a matter of general knowledge and knowledge of the circumstances of the utterance — the situation to which it refers, what has been said earlier, the participants in the discourse, and so on. For our purposes, however, context is the information that is already represented in the model of the discourse.

It is this information that determines how we use the representation of meaning in constructing a model. The meaning of a premise and the existing set of models are used to determine which of the following procedures should be carried out:

Starting a model ab initio
Adding information to a model
Combining models in terms of a common referent
Verifying the premise
Searching for alternative models.

For example, if a sentence does not refer to any items in any existing models, the program uses the meaning of the sentence to start a new model. This model represents the situation referred to by the premise (its extension).

Human beings can draw parsimonious conclusions for themselves; most automated reasoning programs cannot. The task is intricate and intractable but important because in the propositional calculus, it is equivalent to the simplification of an electronic circuit built up from Boolean logic-gates. The standard algorithm for this task (the prime implicant method devised by McClusky 1956, and Quine 1956) is restricted to the connectives "not," "and," and "or." We have implemented a new algorithm based on models (Johnson-Laird 1980b). It can outperform the prime implicant method because it uses the full set of connectives and is guaranteed in principle to find a conclusion as parsimonious as possible. The set of models is recursively divided into pairs of partitions, with the recursion ending when a partition contains only pairs of atomic propositions. The descriptions of these partitions can then be assembled in a way that yields a maximally parsimonious description.

When a program searches for an alternative model, it is at liberty to undo two sorts of information — arbitrary
decisions and default values – and to insert instead some other, specified, value. This possibility confers on a model-based reasoning system the power to make both valid deductions and nonmonotonic inferences. The program maintains a model until, and unless, it conflicts with an assertion. At that point, the model is revised so as to try to satisfy all the assertions in the discourse. If the attempt fails, the current assertion genuinely conflicts with the earlier information built into the model. This process is complementary to deduction, where a search is made for a model that falsifies a conclusion. Such methods are limited to everyday discourse where only a finite set of alternative models needs to be constructed – a particular model is a representative sample and can always be revised so as to satisfy any truly consistent discourse. The moral is that an excellent method for maintaining consistency, whether in a program or a brain, is to work directly with models.

10. Thinking, rationality and models

The whole of our book is one long argument, so we end it with a recapitulation of its principal points. We then consider some consequences of the model theory for the acquisition of deductive competence, for other sorts of thought, and for the debate over whether the concept of rationality is universal or relative to particular cultures.

The acquisition of deductive competence is profoundly puzzling for theories based on formal rules: how could children who know no logic acquire formal rules for valid reasoning? In our view, what has to be acquired is a capacity to build models of the world, either directly by perception or indirectly by understanding language, and a capacity to search for alternative models (see Russell 1987). The acquisition of these abilities, we argue, is less problematic than the acquisition of formal rules.

In daily life, people often lack sufficient information to make valid deductions. They are forced to make plausible inferences that go beyond the semantic information in the premises. Consider the following premises:

The old man was bitten by a poisonous snake. There was no known antidote available.

When subjects were asked what happened, they replied that the old man died (unpublished experiments carried out in collaboration with Tony Anderson of the University of Strathclyde). But when asked whether there were any other possibilities, they could envisage some alternatives. Everyday inferences are plainly not deductively closed: arguments can in theory produce ever more banal possibilities, for example, the old man was kept alive long enough for someone to invent an antidote. At no point can a stage be reached in which all the alternative possibilities have been eliminated. Hence, inferences of this sort lie outside deduction (Collins & Mieghalski 1989). Likewise, the arguments that people construct in favour of particular propositions are not deductively valid: in collaboration with Mark Kean, of University College Dublin, we have confirmed this intuition (common to many investigators of informal reasoning) by asking subjects to construct an argument for such propositions as:

The government should subsidize ballet.

Logic is not the primary guide to this process. Hence, many of the inferences of daily life cannot be accounted for by formal rules that are deductively valid. The process appears to be one in which individuals add new semantic information to their models.

In summary, as Kenneth Craik (1943) proposed long ago, thinking is the manipulation of models. Our research corroborates the claim for deduction, but other modes of thought – induction, analogy, creative problem solving, decision making, and the generation of new ideas – are likely to be based on models too. It may be an egregious error to assume that the representations underlying these other modes of thought take the form of propositional representations or semantic networks, which have structures that are very different from those of mental models.

Do the criteria for rationality – whatever they may be – apply across all cultures, or are the criteria themselves relative to a culture? This question has perplexed all those who have thought about it and has split them into two opposing camps: rationalists, such as Hollis (1970), argue for a core of rational cognitive principles common to all human societies; relativists, such as Barnes and Bloor (1982), argue for purely “local” criteria of rationality, the incommensurability of the beliefs of different groups (even those of scientists of different theoretical persuasions), and the radical unreconstructability of such beliefs from one language to another. If relativism is right, then the principles of deduction differ from one society to another, and perhaps from one epoch to another – as certain historians have argued (see Burke 1958). Hence, psychological studies of deduction are at best parochial interest. Most of the debate, however, has been conducted with scant regard for psychological evidence.

We argue that the model theory provides a way to resolve the controversy. There is a central core of rationality, which appears to be common to all human societies. It is the semantic principle of validity: an argument is valid if there is no way in which its premises could be true and its conclusion false. A corollary of the principle is that certain forms of argument are valid, and these forms can be specified by formal rules of inference. It is a gross mistake, however, to suppose that these rules are per se cognitive universals. Rationality is problematical if it is supposed to be founded on rules. This foundation makes relativism attractive because systematic error is hard to explain, unless one abandons rationality in favour of alternative, and illicit, rules of inference (as some theorists, such as Jackendoff 1988, seem prepared to do).

Finally, we present a critique of mental models. Adherents of formal rules have, not surprisingly, made many criticisms of the theory. Our work has already answered the charge that the theory is empirically inadequate (Braine et al. 1984; Evans 1987) – that it does not apply to propositional reasoning, or to Wason’s selection task, or to inferences in general. We therefore reply to the other objections, which we divide into three categories. The metaphysical criticisms concern the theory’s violation of the tenet that “cognitive psychology has to do without semantic notions like truth and reference” (Oden 1987, Rips 1986), and the claim that models are unnecessary because theorists can rely solely on propositional representations (Plyshyn 1981), neural events (Churchland 1986), or some other reductive format. The methodologi-
11. Conclusions

The puzzle of deductive reasoning may seem parochial for anyone not embroiled in it. We can imagine such a reader thinking: "Deduction is a small and perhaps artificial domain, hence does it really matter whether people reason by manipulating formal rules?" Indeed, the puzzle may seem more serious than the failure to settle a border dispute among warring theories of deduction: controversies in cognitive science may be beyond the scope of empirical resolution.

Our book raises two main questions: how to characterize human deductive competence and how to describe its underlying mechanism. It proposes an account of both questions. If formal rules of inference were in the mind, then the development of logic as an intellectual discipline would be largely a matter of externalizing these principles. And if "psychology" were correct, then logic would be merely the systematization of the natural principles of thought. We reject both these doctrines. Nonformal logic exists in the heads of anyone other than logicians. The principles of thought are not formal rules of inference.

Why then have so many theorists in so many disciplines advocated formal rule theories? One reason is the weight of tradition; another is the greater accessibility of formal accounts of logic (see also McDermott 1987). We have tried to show that deduction can be carried out by other means, and that these means are more plausible psychologically. Let us sum up the case for the model theory. Although the mechanism that enables individuals to make deductions is not available to introspection, experimental evidence shows that the content of premises with the same logical form can have a decisive effect on what conclusions people draw. The late Jean Piaget discovered this effect, and introduced a clause in small print — the "horizonal decalesce." Essentially, the deduction of the phenomenon — to try to sweep it away. Yet the phenomenon is inimical to formal theories of inference. The evidence also shows that when people reason they are concerned about meaning and truth. They are influenced by what they believe to be true, which affects both the conclusions they formulate for themselves and their evaluation of given conclusions. When they draw their own conclusions, they maintain the semantic information from the premises and treat conclusions that throw it away as improper. And, without exception, the results of our experiments corroborate the model theory's predictions about propositional, relational, quantification, and metalinguistic reasoning. Easy deductions call for one explicit model only; difficult deductions call for more than one explicit model, and erroneous conclusions usually correspond to only one model of the premises.

We claim that the model theory accounts for all the robust findings about deductive reasoning and that it successfully predicts novel phenomena. We conclude that logically untrained individuals normally reason by manipulating mental models; we acknowledge that they are able to develop rudimentary formal rules by reflecting on their own performance (Galotti et al. 1986), but such rules are neither complete nor part of their normal reasoning mechanism. There are many uncertainties, gaps, and perhaps downright flaws in the theory of mental models.

Yet, we are convinced of the truth of its broad view — at least to the degree that anyone ought to be committed to a theory. The search for counterexamples can be carried out by constructing alternative models. The method makes an excellent system for computer reasoning. The evidence suggests that it is the mainspring of human reasoning.

NOTE
1. The same difficulty of keeping track of disjunctive models may underlie other deductive puzzles (see Grof & Newstead 1989).