BEHAVIORAL AND BRAIN SCIENCES (1993) 16, 323–380
*Printed in the United States of America*

# 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:** *rmbyrne@vax1.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 resolves long-standing puzzles about reasoning, including how nonmonotonic 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 metalogical inferences based on assertions about the true and the false. Where the two theories make opposite predictions, the results confirm the model theory and run counter to 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; nonmonotonic reasoning; quantifiers; rationality; reasoning; syllogisms; theorem proving

> I'm thirsty, he said. I have sevenpence. Therefore I buy a pint. . . .
> The conclusion of your syllogism, I said lightly, is fallacious, being based on licensed premises.
>
> Flann O'Brien. *At Swim-two-Birds* (1939, p. 20)

> 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, "Here is a gentleman of a medical 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 in 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 that you came from Afghanistan, and you were astonished.
>
> Arthur Conan Doyle. *A Study in Scarlet* (1892, p. 24)

## 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 murkier 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 (1955/1969), Hintikka (1955), 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-

 0140-525X/93 $5.00+.00

ophers. It has also been advocated by some psychologists (e.g., Henle 1978). Others take a much darker view about logical competence and propose theories that render human beings intrinsically irrational (e.g., Erickson 1974). Our view is that people are rational in principle, but fallible in practice. Even though they are not normally 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 metalogical intuitions are important because they prepare 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 validly 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 banal, such as the conjunction of a premise with itself, and no sane individual, 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 maintain the information conveyed by 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.

No amount of data, of course, can pick out one theory against all comers, 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:

If x is a psychologist then x is an experimenter

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

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 1975; Macnamara 1986; Osherson 1975; Pollock 1989; Reiter 1973; Rips 1983; Sperber & Wilson 1986). Each connective, such as "if," "and," and "or," has its own rules. 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:

If there is a triangle then there is a circle.
There is a triangle.
Therefore, there is a circle.

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

If there is a triangle then there is a circle.
There is not a circle.
Therefore, there is not a triangle.

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

If p then q
p
∴ q

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

| | |
|---|---|
| If p then q | |
| not-q | |
| p | (by hypothesis) |
| ∴ q | (by *modus ponens*) |
| ∴ q and not-q | (by conjunction) |
| ∴ 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; Rips 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 matter 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-

cause of the limited capacity of working memory. An initial understanding of a conditional, such as:

If there is a triangle then there is a circle

yields a model of the possible state of affairs in which the antecedent is true and an alternative model of an implicit state of affairs, which in these diagrams we shall represent by three dots:

△ ○
. . .

The implicit model may be fleshed out later with an explicit content. Meanwhile, these two models are neutral between a conditional and a biconditional interpretation (i.e., if, and only if, there is a triangle, then there is a circle). The first step towards the conditional interpretation is to represent the antecedent as *exhausted* as shown by the brackets:

[△] ○
. . .

Exhausted elements cannot occur in models created by fleshing out the content of an implicit model. Hence, in this case, if the implicit model is rendered explicit, it cannot contain a triangle but it can contain a circle. The biconditional interpretation calls for both antecedent and consequent to be exhausted:

[△] [○]
. . .

After the construction of models of the premises, the next step is to formulate a putative conclusion, if none is provided by a helpful experimenter. Because the process is based on models of the premises, it naturally maintains their semantic information. No conclusion can be drawn in some cases unless an implicit model has been fleshed out explicitly. If there is a conclusion, the final step is to search for alternative models that might refute it. The conclusion is valid if there are no such counterexamples. According to this theory, the difficulty of a deduction depends on two principal factors: whether implicit information has to be made explicit and whether the deduction depends on the construction of more than one model. We examined these predictions in all the major domains of deduction, and we now turn to a synopsis of our results.

## 3. Propositional reasoning

The model theory explains all the robust phenomena of propositional reasoning. *Modus ponens* is easier than *modus tollens* because of the explicit information in initial models of conditionals. The conditional:

If there is a circle then there is a triangle

elicits one explicit model and one implicit model:

[○] △
. . .

The premise for *modus ponens:*

There is a circle

eliminates the implicit model, and so the conclusion is immediately forthcoming from the remaining explicit model:

There is a triangle.

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:

[○] [△]
¬[○] [△]
¬[○] ¬[△]

where "¬" 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:

[○] △
¬○ ¬[△]
. . .

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:

| | | |
|---|---|---|
| [c] | [t] | [s] |
| [c] | [h] | [s] |
| [c] | [h] | |
| [c] | | [s] |
| | [t] | [s] |

where "c" denotes Linda in Cannes, "t" denotes Mary in Tripoli, "h" denotes Mary in Havana, and "s" denotes Cathy in Sofia.[1] 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, [c] [t] [s] and [c] [s], and it is invalid because other models falsify it, for example, [c] [h]. Rule theories, however, have yet to lead to the discovery of novel phenomena; adherents have fitted their theories to data from variegated 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 fully 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 tigers 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 1976) dovetails with this account.

Models can be interrelated by a common referent 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; 1983; Wason & Johnson-Laird 1972). In the original version of the task, four cards are put in front of a subject, bearing on their uppermost faces 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 3 card. Yet, the combination of an A with a 3 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, those 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 content-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 1984). It follows that pragmatic reasoning schemas cannot be the whole story. 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 fleshing out of the models of the conditional with explicit representations of the false consequent 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:

[A] 2
. . .

or:

[A] [2]
. . .

and so people tend to select only the A card, or the A and the 2 card. The model theory is thus compatible with Evan's matching bias on the assumption that negation leads to fleshing out the models with the state of affairs that is denied (Wason 1965). Likewise, experience with the rule about beer drinking helps to flesh out the models with more explicit information:

[drinking beer] over 18
¬ drinking beer [¬ over 18]
. . .

and so subjects will now tend to select the card corresponding to the negated consequent.

Cosmides (1989) has argued that insight into the selection task depends on the evolution of a specific inferential module concerned with violations of social contracts. She shows that a background story eliciting such ideas can lead subjects to a surprising selection: they choose instances corresponding to not-p and q for a conditional rule of the form, if p then q. In the context of the story, the rule:

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 meaning:

A man may eat cassava root only if he has a tattoo.

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

[eats cassava] tattoo
¬ eats cassava [¬ tattoo]
. . .

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:

For any x, y, z, if x is in the same place as y, and y is in the same place as z, then x is in the same place as z.

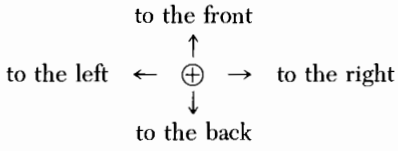
The model theory does not need such postulates. Their work is done by a representation of the *meaning* of the relation, that is, its contribution to truth conditions. One advantage is that the logical properties of relations emerge from their meanings. It is then easy to see why certain relations have 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:

Matthew is to the right of Mark.
Mark is to the right of John.

lead naturally to a transitive conclusion:

∴ Matthew is to the right of John

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:




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; Huttenlocher 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:

$$\text{Left}(x, y) \,\&\, \text{Front}(z, x) \rightarrow \text{Left}(\text{front}(z, x), y)$$

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 - B | B - A | A - B | B - A |
| B - C | C - B | C - B | B - C |

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

Given that each premise 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 & Huttenlocher, reported in Johnson-Laird 1975). One result was the discovery of a very robust "figural" effect.

In general, a syllogism in the figure:

A - B
B - C

tends to elicit conclusions of the form:

A - C

whereas a syllogism in the figure:

B - A
C - B

tends to elicit conclusions of the form:

C - A

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 reordering of information in a premise and a reordering of the premises themselves to bring the two occurrences of the middle term into temporal contiguity (Johnson-Laird & Bara 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 1969; Woodworth & Sells 1935). This notion of an "atmosphere" effect continues to exert its influence on recent theories (e.g., Madruga 1984; 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 1981), 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 untutored 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 (interrelating 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]  b
[a]  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]   [b]
[a]   [b]
[¬a]  [b]
[¬a]  [¬b]
```

where "¬" 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     [a]
 b     [a]
[¬b]   ¬a
 . . .
```

The implicit individual can be fleshed out as:

```
 b     ¬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 corroborated the theory (Johnson-Laird & Byrne 1989).

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 them 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

or its equivalent:

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:

| [○] [○] [○] | [△] [△] [△] [+] [+] [+] |

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:

| [○] [○] [○] | [△] [△] [△] + + + |

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:

| [○] [○] [○] + | [△] [△] [△] + + + |

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 Tabossi 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. 1989). 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 *metalogical* reasoning, 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 metalogical 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 1990). 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 three 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 to 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 McCluskey 1956, and Quine 1955) is restricted to the connectives "not," "and," and "or." We have implemented a new algorithm based on models (Johnson-Laird 1990a). 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: reasoners can in theory produce ever more baroque 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 & Michalski 1989). Likewise, the arguments that people construct in favour of particular propositions are not deductively valid: in collaboration with Mark Keane, 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 untranslatability 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 1986). Hence, psychological studies of deduction are at best of 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 only 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 (Pylyshyn 1981), neural events (Churchland 1986), or some other reductive format. The methodologi-

cal criticisms are that the theory is not clear (Goldman 1986), that it relies unfairly on a visual metaphor (Ford 1985), that it is unworkable (Rips 1986), and that it is untestable (anonymous referee). The logical criticisms are that mental models are irrational (anonymous referee), that there is little or no difference between mental models and formal rules of inference (Goldman 1986; Rips 1986), and that mental models are nothing more than the first-order predicate calculus. Here, we will spare readers a synopsis of our replies to these criticisms; we anticipate that we will have to rehearse them again in our Response to the reviewers. [See also Cohen: "Can Human Irrationality be Experimentally Demonstrated?" *BBS* 4(3) 1981 and Kyburg: "Rational Belief" *BBS* 6(2) 1983.]

In our view, the major shortcoming of the mental-model theory is its incompleteness. Consider, for example, the search for alternative models. Reasoners appear to formulate an initial conclusion, which they often abandon in the light of an alternative model. We see signs of the process in the pattern of their errors, in the revisions they make when they are allowed to reevaluate their conclusions, and in their erroneous memories for conclusions they have drawn. But how are the alternatives generated, and how is the search for them terminated? We do not know.

Likewise, how does the mental-model theory fit into a general account of cognitive architecture of the sort, say, that Newell (1990 [see multiple book review in *BBS* 15(3) 1992]) has proposed? Again, we do not know. Our research suggests only two conclusions about such explanatory frameworks. First, metadeduction calls for mental representations that have an explicit symbolic structure, that is, the assessment of the truth-preserving properties of a particular form of argument requires an explicit representation of that form. Other components of deductive processing may depend on low-level processes that use distributed representations in connectionist networks (Rumelhart 1989). Second, mental models are a form of data-structure that plays a central role in the computational architecture of the mind, entering into not only deduction, but also perception (Marr 1982), the comprehension and production of discourse (Garnham 1987; Johnson-Laird 1983), and the representation of beliefs and other intentional contents (McGinn 1989).

## 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, rules with a specific content, or mental models? They probably use all three.* In fact, the puzzle does matter to cognitive science. None of the three theories appeals to unanalyzable or mystical processes; all are sufficiently articulated to be modelled computationally; and all are feasible explanations of an important mental capacity. Abundant experimental and observational data have been gathered since the turn of the century by psychologists, anthropologists, and others. If one still cannot decide among the three theories, then the consequences may be 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 answers to 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 "psychologism" were correct, then logic would be merely the systematization of the natural principles of thought. We reject both these doctrines. No formal 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 "horizontal décalage," essentially a redescription 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 corroborated the model theory's predictions about propositional, relational, quantificational, and metalogical 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 Griggs & Newstead 1982).

# Open Peer Commentary

*Commentary submitted by the qualified professional readership of this journal will be considered for publication in a later issue as Continuing Commentary on this article. Integrative overviews and syntheses are especially encouraged. All page references are to Johnson-Laird & Byrne's* Deduction *unless otherwise indicated.*

## Mental models and tableau logic

Avery D. Andrews

*Department of Linguistics, The Australian National University, Canberra, ACT, 2601 Australia*

**Electronic mail:** *avery.andrews@anu.edu.au*

Johnson-Laird & Byrne (J-L & B) insist on a strong distinction between reasoning based on mental-model building and that based on the use of formal inference rules. In this discussion I will suggest that the distinction might not actually be so sharp, pointing to a substantial resemblance between what needs to be done by model-construction processes on the one hand and what is done by the processes of tableau development on the other, in the "tableau logic" of Beth (1955/1969), Fitting (1983), Hintikka (1955), and Smullyan (1968).[1] Pursuing and exploiting this resemblance might confer the advantages of a well-defined mathematical framework as well as making it easier to derive psychological hypotheses from current logic-based work in formal semantics.

In J-L & B's model of reasoning, an argument is evaluated by first constructing a mental model that supports the premises and then seeing whether it supports the conclusion. If it does, the next step is to explore alternative models that support the premises; if they all support the conclusion, the argument is accepted. Otherwise it is rejected, as a result of being able to imagine a situation where the premises are true but the conclusion is not.

Tableau logic is similarly based on the idea of a search for counterexamples, but with certain differences. Perhaps the most important of these is that in a tableau logic there is a systematic method of searching for counterexamples to an argument (models that satisfy the premises but not the conclusion) that is guaranteed to terminate with failure if the argument is valid (all attempted ways of constructing a counterexample yield a contradiction). My proposal is that mental-model building can be regarded as being a kind of tableau development, but without the guarantees.

The basic intuition behind tableau development is the idea that the truth of complex sentences is systematically related to the truth of simpler ones. For example, if "Mary is sick or Susan is well" is true, then either "Mary is sick" or "Susan is well" is. One thus starts with the premises of the argument and the negation of its conclusion; then one adds to this collection of sentences simpler sentences whose truth is required by, and sufficient to deliver, the truth of the original sentences. If there are multiple possibilities, each is explored independently, constituting a branch of the tableau. So our sentence about Mary and Susan would give rise to two branches, one containing *Susan is well* and the other containing *Mary is sick:*

Susan is well or Mary is sick

Susan is well | Mary is sick

On the other hand, suppose we have the additional premise *nobody is sick*. A suitable tableau development rule for *nobody* would require that any name *c* appearing in a tableau branch with this *nobody*-sentence, *c is not sick*, also appear in that branch. Hence both branches of the tableau must contain *Mary is not sick* and *Susan is not sick*, so the right branch will be excluded as contradictory, leaving the left as the sole consistent branch that supports the original premises. Note that while some kinds of assertions can be used once and then forgotten about, others, such as universals and negative existentials, have to be maintained as "standing instructions."

J-L & B's notations diverge in certain respects from what one would expect to find in a tableau development system. For example, they use a square-bracket notation to indicate that whatever is in the brackets is exhaustively represented. In the model they propose for "Either there is a circle or else there is a triangle, but not both," a circle appears on one line and a triangle on another, each line representing an alternative mental model, equivalent, I propose, to a tableau-branch:

[○]

[△]

What the brackets indicate is the mutual exclusivity of this disjunction: Something appearing in square brackets is supposed to be "exhaustively represented," not appearing elsewhere in the set of models. So the brackets around the circle are intended to keep circles from appearing on the triangle line and vice versa, thus preventing the sentence from describing a situation in which there is both a circle and a triangle.

The difficulty with this notation is that its exact formal interpretation is not at all clear; and, concomitantly, neither is its meaning in various kinds of complex situations. Suppose, for example, that our sentence was "either there is a circle or a triangle, or a triangle and a square, but not both." It is unclear to me how one ought to deploy the brackets to deal with a case of this nature and what the rules for doing so would be.

The concepts of tableau-development provide an alternative approach to mutual exclusivity: The disjunction provokes the construction of two branches, one asserting the existence of a triangle, the other that of a circle. Additional information might motivate the addition of additional material to either or both branches. But the negative trailer, "but not both," also specifies a condition that all branches must satisfy (by virtue, of course, of some highly nontrivial linguistic principles of ellipsis). I would conjecture that the actual meaning of "or" in English is nonexclusive, with exclusion, when it is understood, being derived from standing assumptions that are part of the context.

Not all of J-L & B's models can be naturally construed as sets of atomic sentences: Extensive use is also made of "spatial models" that appear to be essentially a sort of image (a representation with "pixels" corresponding to regions of space, with a direct representation of directional and adjacency relations between pixels). What I suggest here is that sets of atomic sentences might serve as an interface between language and these spatial models. This would achieve an integration between two senses of "mental model" that seem to exist in an unacknowledged cohabitation in *Deduction*.

NOTE

1. The existence of this work is noted in *Deduction* (p. 16), but J-L & B do not explore its relationships to their own ideas.

## Getting down to cases

Kent Bach

*Department of Philosophy, San Francisco State University, San Francisco, CA 94132*

**Electronic mail:** *kbach@sfsuvax1.sfsu.edu*

Lewis Carroll (1895) posed a paradox of deduction: An argument's premises P cannot be shown to yield the conclusion C unless it is assumed that if P, then C. But if that assumption is included as a premise in the argument, then it must be assumed further that if [P and if P, then C], then C. And so on, *ad infinitum*. So how can deduction get off the ground? The usual

answer is that what is needed is a rule of inference, not an additional premise. A different answer is to reject the assumption that deduction involves proof. Instead of showing that a conclusion follows from the premises, consider the possibilities. In particular, first make sure the conclusion is compatible with the premises and then check to see that no contrary conclusion is compatible too. Now some would call this method proof by cases. And Lewis Carroll might have wondered whether such a method could qualify as deduction unless it includes this assumption: The conclusion follows from the premises just in case it is compatible with the premises and no contrary conclusion is compatible too. It would be awkward (they are averse to rules) for proponents of this method to claim that no such assumption is needed so long as there is a rule to that effect.

Be that as it may, formal logic fares badly when adapted to the psychology of deductive reasoning. For one thing, most valid arguments are not *formally* valid, and the additional assumptions (meaning postulates) needed to make them formally valid are not psychologically realistic. As Johnson-Laird & Byrne (J-L & B) show, rule theories (R) based on formal logic are descriptively inadequate or else *ad hoc*. This should come as no surprise to anyone who has read Harman (1986, Ch. 1). R is bound to be a nonstarter as a theory of reasoning, because logic is the theory of validity, not of reasoning, and reasoning in real life (where arguments are generally not labeled as such) is never just a matter of assessing the validity of arguments. Given a deductively valid argument, you can reject one of its premises (this is what happens in what J-L & B misleadingly describe as the "suppression" *Deduction*, [p. 83] of valid deductions) rather than accept its conclusion, and which choice you make is not a matter of deduction. Besides, although rules of inference are involved in proof, proving (or disproving) a putative theorem is one thing and, as logicians and mathematicians know all too well, thinking of a theorem to prove is quite another.

J-L & B pit their model theory (M) against rule theories, which include those with content-specific rules designed to handle various content-sensitive reasoning phenomena. But M and R, unless defined so broadly as to leave nothing out between them, are not the only possibilities. My model of the rival theories looks something like this:

R

M

. . .

The ellipsis implicitly represents alternatives not yet dreamt up. Having read *Deduction*, I've ruled out R, but I'm not yet ready to opt for M, even though I haven't dreamt up an alternative. While allowing that countless alternatives are possible, J-L & B take R to be M's only serious competition. R may be the only existing rival to M, but that doesn't make evidence against one automatically count in favor of the other. J-L & B provide no direct evidence that people actually use the proposed models but claim only (no small claim) that M is consistent with the evidence. But so might be a whole host of other theories. For example, M might not be the only one that "implies that people search for counterexamples" (p. 39). What worries me here, however, is that J-L & B have not formulated M specifically enough to make clear what it does and does not include and what is and is not essential to it. In the hope of finding this out, I will pose several sets of questions, which raise issues at distinct levels of explanation.

A theoretical account of phenomena at one level can be compatible with different accounts at the next. Noting the difference between explaining *what* deductive competence involves and *how* it operates (p. 17), J-L & B cite Marr's (1982) distinction between the computational and the algorithmic levels, his levels 1 and 2 (his physical realization level 3 is not relevant here). Several intermediate levels are needed as well. Peacocke (1986) has proposed an informational level 1.5, and I suggest a further intermediate level, a representational level between the informational and algorithmic (1.75, if you insist). One concerns what information feeds into an algorithm and the other concerns how (in what format) this input is represented. Here let us agree with J-L & B that at the computational level, the level at which deductive competence is characterized, people are able to make valid deductions, though they often miss valid ones and make invalid ones, and in real life draw new or useful conclusions (pp. 18–22). Also, let's assume that what counts as a rule theory is significantly restricted, so that not every theory with an algorithmic level counts as a rule theory just because it can be implemented in a program. Now, given these assumptions and the above levels of explanation, there are several kinds of questions, answers to which would give M much more specificity.

***1. Encoding.*** Students of logic and language are well aware of the notorious problems of formalizing various features of natural language. M has an analogous problem: J-L & B often mention how a sentence "calls for," "leads to," or "yields" the initial construction of a model, but precisely how does a sentence get encoded into a model? How is it determined which information does or does not go into the model, especially if "mental models are remote from the structure of sentences" (p. 212)?

***2. Representation.*** Until J-L & B constrain the notion of model, their claims about how people reason, for example, by searching for a counterexample, are not specifically model-theoretic. What forms of representation count as models? For example, do pictorial (Bach 1970) and other systems of graphical representation (Goodman 1968) qualify? How is information in a set of models tracked, so that the information contained in each can be utilized in working with the whole set? For example, how does the exhaustiveness explicitly represented by square brackets in one model constrain what goes into later models? Finally, what is the difference, in M, between explicit and inexplicit information (Cummins 1986)?

More specifically, how is it determined which features of a given model do the representing and which states of affairs are thereby represented? For example, spatial relations in a model often represent spatial relations in what is modeled, but often they do not. What determines when they do and when they do not? In general, what ensures that all relevant information enters without any irrelevant information also creeping in (recall Berkeley's worry, in criticizing Locke's theory of abstract ideas, that the general idea of a triangle must, because triangles come in different forms, be "all and none of these at once"; 1710/1965, p. 13)? J-L & B appreciate such questions when doing programming for reasoning (Ch. 9), but otherwise their encoding and representational schemes seem rather improvisional in character.

***3. Strategy.*** What drives the search for alternative models? How does one know when no more models are needed? J-L & B appreciate these questions, as when discussing people's deductive limitations and the nature of default reasoning, but they do not really answer them. However, I applaud J-L & B's contention that R cannot accommodate default reasoning, for in my view (Bach 1984) jumping to conclusions is justified only insofar as one can rely on one's ability to think of (not the same as to deduce) reasons to the contrary when they are worth considering.

***4. Scope.*** J-L & B assert that "propositional, relational, and [standard] quantificational reasoning exhaust the main sorts of deduction" (p. 3). These areas offer plenty to worry about, but there is also deduction involving nonstandard quantifiers (e.g., *few, several, many, most*), quantitative reasoning, modal reasoning, and conceptual reasoning. How well does M work for them? What additional kinds of models do they require?

***5. Diversity.*** J-L & B often speak of one deduction problem being "easier" than another, as indicated by different aggregate error rates. But perhaps the data indicate something more

complex. Could some "hard" problems be easy for some people? Why do some people solve problems much more quickly and accurately than others? Do they use different models and strategies? Testing for individual differences in speed, accuracy, and characteristic errors might provide some answers.

Although some of these questions raise worries similar to Rips's (1986), I do not share his skepticism about M, much less his enthusiasm for R. For one thing, deduction is not an autonomous domain and the mental-model model is more plausible for thinking generally than is any rule theory. However, I do think that such questions as the above must be answered before M can jell into a clearly defined theory. M needs to make explicit what is involved, at each explanatory level, in deduction by modeling, which, after all, is not just a matter of imaginative proof by cases.

## Toward a developmental theory of mental models

Bruno G. Bara

*Centro di Scienza Cognitiva, Universita' di Torino, 10123 Turin, Italy*

As a long-standing collaborator of Philip Johnson-Laird's, I can be credited with a firm faith in the validity of the mental-model approach to human reasoning. I shall therefore not embarrass Johnson-Laird & Byrne (J-L & B) with praise, but concentrate instead on the points still open to criticism.

The experimental evidence reported by J-L & B is persuasive regarding the greater fruitfulness of model theory relative to logic-in-the-mind approaches; however, one may note that nearly all J-L & B's data refer to adults. Although the authors are undoubtedly aware that their proposal does not cover the developmental aspect of deduction, they still undervalue this problem. In Chapter 10, J-L & B show why the acquisition of deductive competence is puzzling for theories based on formal rules. Their well-founded criticisms lead to the two questions I raise in my commentary:

1. Is there a mental-model theory of the acquisition of deductive competence?

2. Granted that such a theory could be formulated, would it present fewer puzzles than previous rule theories?

The current answer to the first question is plainly no. This accordingly represents a shortcoming of all candidate theories, because any competence shown in the adult system is not fully understood until its development from childhood is explained. The work of Bruner et al. (1966) and Chomsky (1965) attests to the relevance of concept and language acquisition, respectively. This shortcoming grows into a misfortune in this area, thanks to the influence of Piaget (1977), the best known developmental psychologist who devoted the largest part of his work to the processes underlying the development of the child's logical competence until the adolescent level of formal operations is attained. True, Piagetian positions have been continuously attacked, but no convincing alternative has yet been advanced.

Model theory is a natural alternative, provided it offers an adequate coverage of the domain. The fact that Piaget's school still dominates psychologists' view of how deductive abilities are acquired makes it harder to change to a different paradigm for adults. It is like walking from Paris to Amsterdam and being told on your arrival that you have reached Rome. Once the scientific community is presented with a comprehensive – albeit general – view of how model theory explains the transition from infant to adult performance, the evidence against formal logical theories should become much more acceptable. To gain full explanatory power, an ambitious theory has to show how the phenomenon under examination has been caused or constructed; for model theory this means drawing the lines of development of the basic abilities involved in the construction and manipulation of mental models.

The necessity of a developmental model theory having been pointed out, the second question can be faced: What are the difficulties of such a theory. In Chapter 10 J-L & B write about the pillars of the acquisition of deductive competence: the capacity to interpret the world, by perception or by language, and the capacity to search for alternative interpretations. As long as the first capacity is involved, things should be not too problematic: The ability to master language and the processing capacity of working memory surely increase from childhood. I shall concentrate instead on the latter, which the authors consider the common denominator of rationality.

To look for a counterexample, a system needs more basic abilities. These include at least the capacity to construct models, to confront different models among them, and to compare a model with a verbal description. Again, the development of these basic abilities from neonates to adolescents is easy to show. The puzzle starts with the composition of the basic abilities into a more complex capacity that J-L & B call the mainspring of human reasoning.

How is it that such a powerful strategy of thought gets learned? It does not seem to be an inborn characteristic of mankind, such as the capacity to detect similarities and differences between two images. The vast majority of individuals prefer to stick to positive instances, whether they come from an African tribe (see the Azande defenders of the "poison oracle," cited in J-L & B's Ch. 10) or from a Western university (see the experimental subjects in the four-card selection task, analyzed in Ch. 4). Becoming aware of the usefulness of the search for counterexamples is also difficult. Wason and Johnson-Laird (1972) report that 38.2% of subjects resisted any insight about the correct method in the four-card task, even after the explanation was given by the experimenters in a free interview. Persons find it arduous to grasp why they ought to look for negative instances instead of positive ones.

Even in the field of modern science we find the same difficulty. Karl Popper first proposed his epistemology based on falsification in 1934, but despite its striking the world of science as something absolutely innovative, many years had to pass before it prevailed over the more obvious – and less efficacious – verificationist methodology. This suggests that the search for counterexamples is not culturally intuitive, exactly as it is not immediately natural for all individuals. We are now at the core of the problem: J-L & B have shown that the search for counterexamples is the strategy adopted by adults when they perform various kinds of deductive tasks. What is absolutely unclear is how such an effective method develops at the competence level. Awareness of the utility of looking for negative instances is a second step, difficult for both individuals and for cultures.

## Deduction as an example of thinking

Jonathan Baron

*Department of Psychology, University of Pennsylvania, Philadelphia, PA 19104-6196*

**Electronic mail:** *baron@cattell.psych.upenn.edu*

A modest generalization of Johnson-Laird & Byrne's (J-L & B's) mental-model theory of deduction can describe all goal-directed thinking, including decision making, planning, and design (Baron 1988; Galotti 1989): Thinking proceeds by search for possibilities, evidence (both positive and negative), and goals (including subgoals, which can also be partially described possibilities). Goals are criteria for evaluating possibilities in the light of evidence. In decision making, for example, possibilities are options and goals are criteria for evaluating the options.

Applying this framework to deductive thinking usually yields a two-level structure of thinking episodes. A high level derives conclusions from models, and a low level derives models from premises. The goals of the high level are to derive conclusions (possibilities) that are consistent with the premises, not explicitly stated in them, parsimonious, and so forth. At this level, the evidence consists of possible models rather than the premises themselves. An initial model provides evidence that suggests the conclusion, and other models are sought that might refute it. New conclusions must then be sought that are consistent with the models (the evidence) obtained so far as well as with possible models not yet found. For example (pp. 121–22), in "All Bs are As; no Bs are Cs," an initial model (in which no As are Cs) yields "No Cs are As." A second model (in which only some As are Cs) refutes this conclusion, but supports "Some Cs are not As," but a third model (in which all Cs are As) refutes this conclusion too. The derivation of the conclusion "Some As are not Cs" from the three models requires both a search for possible conclusions (possibilities) and a search for evidence in the form of other models, including those already found and those not yet found. The search for possible conclusions is directed not only by the models discovered (the evidence) but also by the goals of consistency, novelty, and parsimony.

The search for new models requires a subepisode of thinking in which the goal is to find a model rather than a conclusion. The evidence for a subepisode consists of the stated premises, which serve both to suggest models and to act as evidence for and against them. This search may be directed by the conclusion at issue as well as by the premises themselves. In particular, the search may be specifically for models that contradict a tentative conclusion.

Errors in deduction can result from (a) failure to find a single possibility (typically yielding "no conclusion" when a conclusion is possible) or (b) failure to find evidence against a possibility, thus falsely concluding that it is correct. Both errors can result from insufficient search. The latter error can result from a tendency to direct the search toward positive evidence for a tentative conclusion rather than toward negative evidence (or toward both kinds). Such biased search seems to occur in content effects, as J-L & B note.

J-L & B suggest that insufficient search results from limited working memory, but working-memory limits can sometimes be overcome by spending extra time, so it is also possible that insufficient search results from lack of motivation or from overconfidence in initial conclusions (Baron et al. 1986; 1991).

Much evidence suggests that misdirected search – toward positive evidence – is a common source of error in thinking. A number of experiments (e.g., Anderson 1982; Arkes et al. 1988; Hoch 1985; Koriat et al. 1980) have found that asking subjects for reasons why an initial conclusion is correct has little effect on various errors (since people seem to search for such evidence anyway), but asking for reasons why a conclusion might be incorrect reduces errors. Beattie and Baron (1988) also found direct evidence for such biased search in some subjects: When asked to test a rule such as "if the card has an A then it has a 3" by specifying the complete content of any cards they would want to know about, these subjects asked only about cards with A and 3. They did not ask about the existence of possible counterexamples such as A and 2. Baron (1991) and Kuhn (1991) have suggested that part of the problem is that some people do not understand the importance of negative evidence. Instruction in redirecting search away from only positive evidence has been part of a variety of efforts to improve thinking (Perkins et al. 1991; Selz 1935). Baron et al. (1986) found that deductive performance improved in some students after a training program that emphasized more extensive search as well as redirected search.

In sum, the mental-model theory is consistent with a view of thinking in which the ability to solve logic problems has no special status. Logic problems are just one kind of problem that can be thought about, along with arithmetic problems, moral dilemmas, and creative tasks. In all of these areas, people search for possibilities and evidence, and often for goals too (especially in creative tasks). And errors result from too little search (which might be justified, given the costs and expected benefits of thinking), and from partiality to initial possibilities (which is more difficult to justify).

Of course, logic does have special status in that it provides a normative model for the evaluation of inferences – as do probability theory and decision theory (Baron 1988). But, in all of these areas, normative models themselves are only rarely useful as prescriptive guides for the conduct of thinking. The mental-model theory, however, suitably generalized, gives us a language in which we can instruct people in how to improve their thinking.

## Everyday reasoning and logical inference

Jon Barwise

*Departments of Mathematics, Philosophy, and Computer Science, Indiana University, Bloomington, IN 47405*
**Electronic mail:** *barwise@phil.indiana.edu*

I suggest that there are two ways to understand the architecture proposed in Johnson-Laird & Byrne's (J-L & B's) *Deduction:* as an architecture involving models and everyday reasoning, or as one involving representations of classes of models and logical inference. These are distinct proposals, but the close connection between the two suggests that logical inference and everyday reasoning might not be as far apart as traditional approaches have made them appear.

First, we must be clear about the difference between everyday reasoning and logical inference. Suppose Claire knows *P:*

> Jason is taller than Melanie.
> Melanie is taller than Max.
> Melanie is over 5′ tall.

Claire might

(a) infer (conclude with good reason) some fact $Q$ implicit in $P$, or
(b) prove beyond a shadow of a doubt that $Q$ is implicit in $P$, and so conclude $Q$.

Does this distinction matter? In both cases, Claire can act on $Q$ without being led astray. Notice, though, that only in the second case can we say that Claire *knows* $Q$. So the distinction must matter. The more common form of inference may well be (a), but logic deals with (b).

The two claims of this commentary are (i) that J-L & B deal with everyday inference, not logical inference, and (ii) that their architecture can be reinterpreted as part of one that deals with full-fledged logical inference. Recall some important points from the book (2 and 3 are the book's main claims):

1. "An argument is valid if there is no way in which its premises could be true and its conclusion false." (p. 209)
2. "The common denominator of rationality is the search for counterexamples: anything else is logical icing on everyday competence." (p. 209)
   "The evidence suggests that [the search for counterexamples] is the mainspring of human reasoning." (p. 215)
3. "The principles of thought are not based on formal rules of inference." (p. 215)

If we read these as being about everyday reasoning, they seem reasonable enough. (1) is common wisdom, and it strongly

suggests (2). The "counterexamples" mentioned in (2) are the "ways" of (1), circumstances where the premises are true and the conclusion false. And if (2) is right, then that strongly suggests (3), at least given the usual systems of "formal rules of inference." But if (2) and (3) are claims about logical inference, the case falls apart.

An argument is invalid if there is at least one counterexample. By contrast, an argument is valid only if every model of the premises is a model of the conclusion, and there are typically infinitely many possible such models. Since "an infinite set is far too big to fit inside anyone's head" (p. 36), the authors assume that "people construct a minimum of models: they try to work with just a single representative sample from the set of possible models, until they are forced to consider alternatives" (p. 36).

Return to our example above and two possible conclusions:

$Q_1$: Jason is over 5′ tall.
$Q_2$: Max is over 5′ tall.

Clearly $Q_1$ follows from $P$ but $Q_2$ does not. Claire can easily conceive of circumstances in which the premises are true and $Q_2$ holds but others in which $Q_2$ fails.

J-L & B's claim (2) is that inference is a search for counterexamples. If a counterexample is found, as in the case of $Q_2$, then Claire knows that the conclusion does not follow. If none are found, as must be the case with $Q_1$, then eventually Claire must give up searching and take it to follow. But since there is no way to search through all possibilities, she has to draw the line someplace. But at just that point Claire loses her claim on logical certainty. She may have reasons to suppose $Q_1$ follows from $P$, but she doesn't *know* it does.

Before rule theorists start to gloat, we note that they face a dual embarrassment. Though they account for how Claire can know a conclusion logically follows (by finding a correct proof), they cannot account for how she can know a conclusion does not follow, as with $Q_2$. The "search till you're exhausted" strategy gives one at best an educated, correct guess that something does not follow. The rule-based and model-based approaches appear to have complementary strengths regarding logical inference. (J-L & B do not exploit their strength here.)

The psychologist might react to all this by saying it is everyday reasoning that is of concern, that the rest is, to quote J-L & B, "logical icing" (p. 209). But a more interesting and constructive response is possible.

The fact that Claire knows both that $Q_1$ follows from $P$, and that $Q_2$ does not, means that she must have come up with something that would count as a proof of $Q_1$, and something that would count as a counterexample to $Q_2$. Such proofs and counterexamples must both be representable "in the head," in some sense. Why suppose they are represented in radically different ways? Let us represent "taller than" by the "left of" relation.

| Jason | Melanie | Max |
|---|---|---|
| | Melanie | 5′ |

This is not a model of $P$; it is a table holding in *every* model of $P$. With this understanding, we can use this representation to infer, in the sense of logical inference, that Jason is over 5′ tall. What matters is that the transitivity of "left-of" is respected by the transitivity of "taller than." With this interpretation we can still treat nonconsequence. Consider the table

| Jason | Melanie | 5′ | Max |
|---|---|---|---|

This represents a nonempty (in fact infinite) class of models of $P$ in which $Q_2$ fails. (A similar table represents models in which $Q_2$ fails, showing that $Q_2$ is not settled one way or the other by $P$.)

To reinterpret the J-L & B architecture, let us construe their structures not as models, but as special representations of classes of models. We must also add new procedures; for example, we would need one that allows us to conclude that a collection of these representations is exhaustive, and another that allows us to conclude that such a representation has at least one model. Motivated in part by Johnson-Laird's (1983) earlier proposals, John Etchemendy and I (Barwise & Etchemendy 1991) have been pursuing such an approach to logical inference. From this perspective, everyday reasoning is a relaxation of logical inference.

## Mental models cannot exclude mental logic and make little sense without it

Martin D. S. Braine
*Department of Psychology, New York University, New York, NY 10003*
**Electronic mail:** *mdsb@xp.psych.nyu.edu*

Johnson-Laird & Byrne's (J-L & B's) entire argument against a mental logic relies on a syllogism based on exclusive disjunction:

Mental models or mental logic, and not both.
Mental models
∴No mental logic.

The "not-both" clause is based on parsimony:

There are mental models.
∴Mental models are people's only method of reasoning, until proved otherwise.

However, J-L & B say little to make the parsimony argument plausible. Why is it implausible to expect that human beings would have evolved a variety of methods of reasoning?

First, mental models cannot suffice for reasoning, given J-L & B's notion of model. J-L & B do not define "model," but they say (p. 212) that models cannot contain variables, that is, they are always concerned with specific instances. That immediately provides reason to doubt that mental models could account for all reasoning. Mathematical statements and mathematical reasoning provide obvious counterexamples. Consider a statement such as, "If the sum of the digits of a number is a multiple of three, then the number is divisible by three." To represent such a statement one needs a variable – something to represent "any number" – clearly a representation containing only specific numbers would not suffice. Note, too, that a proof of the statement would inevitably mention a variable ("any number"); how could one understand the proof and judge its validity if one cannot represent "any number"? Models do play an important role in mathematical reasoning, for example, diagrams in geometry. But even in geometry, there is much more to a proof than the diagram: Principles that contain variables are used, and there is always a final generalization step, from the triangle, circle, or other figure instantiated in the diagram, to any triangle, any circle, and so on – a step that involves a variable (e.g., "any triangle," "any circle").

Variables are needed even for Aristotelian syllogisms (Ch. 6), to account for people's perceptions of necessity. Consider the syllogism:

All the athletes are bakers
All the bakers are canoeists

Suppose that, following J-L & B, we start by considering a model with just two athletes. They turn out to be bakers, and when we incorporate the information in the second premise they turn out to be canoeists too, and we reach the tentative conclusion *All the athletes are canoeists*. The theory now says that careful subjects will try to falsify the conclusion by seeking other models consistent with the premises in which the conclusion does not hold. So, as careful subjects, should we not try out a model with (say)

six athletes, to check that they too turn out to be canoeists? But of course we do not – because we see that no matter how many athletes there are, each athlete has to be a baker by the first premise, and a canoeist by the second premise – but note that *that* is a logical argument which contains a variable ("each athlete") that is not allowed in J-L & B's theory. Thus, J-L & B's theory does not account for people's perception of the necessity of the conclusion without checking other models. (Of course, given some logic, no model is needed to solve that particular syllogism.)

Second, although the theory claims to invoke only specific instances, parts of the theory appear to make tacit use of variables, and even of inference rules. As one example, consider the logic of "[]," the "exhaustivity tag." This ensures that, given a model such as

[a] b
[a] b
. . .

the left-hand column must be fleshed out with [~a], for example,

[a] b
[a] b
[~a] b
[~a] ~b

The meaning of "[]" appears to be given by a tacit inference rule (with variables):

$$\frac{\text{y is an unrepresented entry in a column that contains "[x]"}}{\text{y} = [\sim x]}$$

J-L & B would also need the rule:

$$[\sim\sim x] \equiv [x]$$

Thus, some mental logic is implicit in J-L & B's theory, making it a hybrid theory and blunting the logic/model opposition that J-L & B insist on.

Third, many versions of the mental-logic thesis assume that some logical apparatus is developmentally primitive – part of an innate format for representing declarative knowledge, of a syntax of thought (e.g., Braine 1990; 1992; in press; O'Brien, in press; cf. Fodor 1975). This would be consistent with the very widespread, and perhaps universal appearance in human languages of connectives similar to English *and, or, if,* and negation, in association with the same common inference forms; likewise, one tends to find words for *all* and *each*, and there are other logical elements (e.g., certain modals) that may be universal. A mental logic provides a more natural explanation than mental models of the reason why these particular elements should be so common.

Finally, J-L & B's theory requires the concept of an *unconscious* mental model held in *working memory,* a paradoxical combination: Information-processing theories customarily take the content of working memory as accessible to consciousness (e.g., Ericsson & Simon 1984). Mental models clearly are often accessible to consciousness, as any reader can attest who attempts the spatial relations problems in J-L & B's Chapter 5. (Similarly, in the folk science exemplified in Gentner & Stevens [1983], the models were generally accessible to consciousness.) However, for logic problems, introspection suggests that subjects do not consistently use the kinds of models proposed. Models are usually not reported in propositional problems (in our experience – cf. Braine et al. 1984); for other logic problems, old evidence (Storring, cited in Woodworth 1938), which is consistent with unpublished work of O'Brien and myself, suggests that there is great variation across people and problems in whether models are reported, and in the kind of model reported. I cannot help thinking that the purpose of the unconsciousness postulate is to shield the theory from this obvious kind of evidence. At the very least, J-L & B need to explain the variability in what people report.

In sum, I have argued, first, that mental models cannot suffice for reasoning, given J-L & B's notion of model; second, that some mental logic is implicit in J-L & B's theory, making it more of a hybrid than they allow; third, that the common logical apparatus of human languages argues for a mental logic and is hard to explain using J-L & B's theory; and, finally, that there is a deep problem with J-L & B's rejection of introspective evidence.

## "Semantic procedure" is an oxymoron

Alan Bundy

*Department of Artificial Intelligence, University of Edinburgh, Edinburgh EH1 1HN, Scotland*
**Electronic mail:** *bundy@ed.ac.uk*

***1. Introduction.*** Johnson-Laird & Byrne (J-L & B) are to be congratulated on proposing a new mechanism for deductive inference and for presenting extensive evidence for the psychological validity of this mechanism. I will have no quarrel with this mechanism or with the psychological claims; both deserve attention and further investigation. My argument is against the implied epistemic nature of the new mechanism.

In *Deduction,* the mental-model mechanism is described as a "semantic procedure" (p. 23) and is said to be "compatible with the way in which logicians formulate a semantics for a calculus" (p. 36). Mental models are contrasted strongly with rule-based mechanisms (e.g., pp. 23, 195). The implication, whether intended or not, is that the mental-model mechanism directly addresses the problem of intentionality. A mental-model-based computer program, it seems, would automatically give meaning to computational states.

I will argue that this implication is wrong; mental models have no more to say about intentionality than rule-based mechanisms. The attachment of the adjective "semantic" to a deductive mechanism, or to any computer program, is misleading and confusing. The phrase "semantic procedure" is an oxymoron. Mental-model and rule-based mechanisms differ only in degree and not in kind.

***2. The meanings of "semantics."*** Unfortunately, the issue is clouded because the word "semantics" is used in different ways by different communities. For example, logicians use it to describe a mapping from the expressions of a logical theory to the "meaning" of these expressions. To give a semantics to a logic is to provide this mapping. Tarski provided a semantics for predicate calculus by showing how logical sentences in a theory could be mapped to truth or falsity in a model.

There is an ambiguity about whether these models are aspects of the real world or mathematical theories in their own right. For a semantics to map formulae to their meaning, models should be part of the real world. However, there are several forces encouraging their formalisation as mathematical theories. Formulae in commonsense reasoning are relatively easy to map to the real world. For example, in *loves* (*John, Mary*) the constants *John* and *Mary* map to specific individuals John and Mary, *loves* to the relationship of loving and *loves* (*john, mary*) to the assertion that John loves Mary. Mathematical formulae, for example, 2 + 2 = 4, are harder to map to the real world because the coherence of the mapping presupposes a platonic commitment to the existence of 2, 4, and so on. Couple this with the natural tendency of mathematicians to formalise, and it becomes easier for them to regard models as mathematical theories of sets of objects on which functions and relations are defined. The sense of "semantics" in which it assigns "meaning" is then lost.

Linguists generally use "semantics" to describe, not the mapping to a meaning, but the meaning itself. A semantic representation of a natural language sentence is contrasted with

the syntactic representation. The syntactic representation is the original string of words or a parse tree with these words labelling the leaves. The semantic representation must capture not this grammatical structure but its content. Confusingly, this is usually done by a logical formula; so the linguist's semantics is the logician's syntax!

Computer scientists use the word "semantics" to describe the mapping from a programming language to a mathematical theory. Ironically, this turns the logician's usage on its head. Logical semantics translates a mathematical formula into a program for calculating a truth value; computer science semantics translates a program into a mathematical formula.

Because of their remark on p. 36 of *Deduction* (see para. 1 above), I will assume that J-L & B intend the word "semantics" in the logician's sense. I assume that their mental models are based on Tarski's models of logical theories; that their deductive mechanism is an attempt to reason in the model theory in contrast to rule-based mechanisms that reason in the proof theory. I claim that it is not possible to do this.

***3. Is semantic reasoning possible?*** If we regard Tarskian models as part of the real world, then reasoning with them would entail physically manipulating the real world. This has limited utility. It is not possible to conduct forward planning, hypothetical reasoning, counterfactual reasoning or abstract reasoning by manipulating the current world state. We must reason by manipulating an internal representation of the world.

At this point the problems of intentionality emerge, that is, we need a semantics to map this internal representation onto its meaning. This remains true even if the internal representation is based on a Tarskian model. Calling the manipulation procedure "semantic" does not affect the situation.

Basing a computational reasoning mechanism on Tarskian models presents problems for a finite computer. For example, some models have an infinite domain of objects. Some reasoning involves proving that an infinite collection of objects has a property. Some reasoning involves the representation and use of incomplete or vague information. These problems are solved in rule-based mechanisms by the use of quantifiers, variables, disjunction, and so forth. Some equivalent device is needed in model-based reasoners if they are to have the same reasoning power. J-L & B use such devices in their mental-model mechanism. For example, infinite numbers of objects are represented by a finite number of tokens; incomplete information is represented by having alternative models to cover the range of possibilities.

***4. Are rule- and model-based reasoners different in kind?*** One paradigmatic example of a rule-based deductive system is a resolution-based theorem prover. The rules are formulae of predicate calculus in clausal form representing the axioms of the theory and the negation of the conjecture. The conjecture is proved by *reductio ad absurdum;* the clauses are "resolved" together, usually exhaustively, until the empty clause is derived.

However, resolution can also be viewed as a systematic attempt to check that none of the models of the theory provide a counterexample to the conjecture. The fact that resolution can be viewed in this way goes back to a metalogical theorem of Herbrand's. If the attempt to prove the conjecture fails after a finite search then a counterexample to the conjecture can be read off automatically from the failed attempt. Thus resolution can be viewed both as a rule-based and as a model-based mechanism!

This potential duality was brought home to me forcibly as a result of my first foray into automatic theorem proving. I built a model-based theorem prover for arithmetic called SUMS (Bundy 1973). Its model consisted of a representation of the "real line" as used by mathematicians in informal blackboard arguments. The hypotheses of the theorem were represented by placing points in appropriate positions on this "real line" and the conclusion was then read off from the model.

As I tried to get SUMS to prove harder and harder theorems, this simple idea became more and more elaborate. For example, consequences of the original hypotheses had to be propagated around the model before the conclusion could be read off. The natural propagation mechanism was forward-chaining with rules. After a while I realised that I had just built yet another rule-based mechanism. SUMS was now similar to a standard semantic tableau prover with a bottom-up search strategy. SUMS' progression from model-based to rule-based was incremental. There was no point at which the nature of its reasoning dramatically changed in kind.

***5. Conclusion.*** I have argued that there is no difference in kind between the mental-model deduction mechanism of J-L & B and rule-based mechanisms. Indeed, it is possible to view many deduction mechanisms as simultaneously of both types. The issue of intentionality arises with both types of mechanism, and is not finessed by the use of a model-based approach. To the best of my knowledge J-L & B make no claim to the contrary. However, others may erroneously draw that conclusion from the free use of words like "semantics," "model," and so forth. For this reason I recommend that the word "semantics" be used with extreme caution. It is a highly ambiguous term and has great potential to mislead.

None of this detracts from Johnson-Laird & Byrne's significant contribution in defining a new deduction mechanism and providing evidence for its psychological validity.

ACKNOWLEDGMENT

The work reported in this commentary was supported by a SERC Senior Fellowship to the author.

## Mental models and nonmonotonic reasoning

Nick Chater

*Department of Psychology, University of Edinburgh, Edinburgh EH8 9JZ, Scotland*

**Electronic mail:** *nicholas@cogsci.ed.ac.uk*

Johnson-Laird & Byrne (J-L & B) are equivocal concerning the scope of mental-model theory. On the one hand, they are careful to note that mental models are aimed primarily at explaining deduction, although commonsense inference is not deductive in character. On the other hand, they contend that mental-model theory solves the problem of nonmonotonic reasoning, which is *not* deductive and is characteristic of commonsense inference. This equivocation requires clarification: An account of deductive reasoning casts light on a fascinating if rather arcane human ability; an account of nonmonotonic inference in general would be little short of a theory of thinking. It is not clear, therefore, exactly what J-L & B see as the domain of the mental-model account. I shall argue that mental-model theory does not in fact address the problem of understanding commonsense nonmonotonic reasoning, still less provide a solution to it.

Everyday, commonsense reasoning may be conceived of as a species of inference to the best explanation: It involves inferring from given information to what best explains and is best explained by that information (Fodor 1983). Such inference is nonmonotonic, because the addition of new information can invalidate what were previously plausible conclusions. So, for example, the plausible inference from hearing the sound of purring behind the door to the conclusion that the cat is trapped in the cellar is immediately overridden if I catch sight of the cat in the garden. The premise on which the inference is based, the purring, need not be withdrawn, although another explanation for this fact may be sought. By contrast, in monotonic reasoning, the conclusion of a valid argument can only be challenged if one of its premises is false.

Providing an account of inference to the best explanation is very difficult. Inference to the best explanation encompasses

theory confirmation in science, the problem of inferring the scientific theory which best fits the available data. Understanding inference to the best explanation also presupposes a solution to the notorious frame problem (McCarthy & Hayes 1969). Indeed the problem of nonmonotonic reasoning and the frame problem really constitute the same problem looked at from different points of view. Where nonmonotonic reasoning concerns which conclusions *should* be revised when new information is added, the frame problem involves deciding which conclusions *need not* be revised.

Nonmonotonic reasoning has proved resistant to a vast number of extremely ingenious proposals within artificial intelligence. The principal line of attack has been to attempt to provide a *nonmonotonic* logic which captures commonsense, rather than deductive, inferences (e.g., Hanks & McDermott 1987; McCarthy 1980; Reiter 1980; Shoham 1987). These logics have been dogged by two serious problems (McDermott 1987; Oaksford & Chater 1991). First, it has not been possible to capture the nonmonotonic inferences that people routinely draw (specifically, most computational methods tend only to draw extremely weak conclusions). Second, these methods have been plagued by computational intractability due to explicit or implicit reliance on consistency checking. J-L & B's treatment of nonmonotonic reasoning addresses neither of these difficulties, but I shall concentrate on the first and most fundamental.

Throughout their book J-L & B give examples of how mental models can be constructed and checked to generate inferences licensed by a standard logic (typically, propositional or predicate calculus). This is not surprising, because they are primarily concerned with deductive reasoning. The rules governing the building of models and searching for alternative models ensure that the notion of the validity of the logic concerned is respected. A mental-model account of nonmonotonic reasoning would require similar rules for model construction and search; but how could these rules be chosen? By analogy with the deductive case, we might expect these constraints to respect the notion of the validity of some nonmonotonic logic. J-L & B rightly do not attempt to follow this line, since there is no appropriate nonmonotonic logic to which appeal can be made. Yet, without an underlying logic of some kind, the processes of building mental models will be without any constraint or justification.

There is also a more specific worry. Inference on the basis of a failure to find countermodels, which is at the heart of the mental-model account, appears inapplicable to nonmonotonic reasoning, because, by definition, such countermodels invariably exist: If there were no model in which the premises are true and the conclusion false, the reasoning would be valid according to a standard, deductive, monotonic logic, and would not be an instance of nonmonotonic reasoning at all. Furthermore, not only do countermodels exist, but people find it easy to generate these countermodels on demand. For example, returning to the inference that the cat is in the cellar on the basis of hearing a purring sound, it is easy to generate a host of countermodels in which this inference is not correct. For example, the sound may be produced by a different cat, a tape recorder, a person doing a cat imitation, the wind, and so on. Quite generally, it is easy to find countermodels for everyday nonmonotonic inferences. This means that if nonmonotonic inference proceeded from a failed search for countermodels, no nonmonotonic inferences would be drawn at all! Notice that the mental-model theorist cannot argue that only plausible models are constructed, or that the single most plausible model is chosen, because the problem of finding such models is simply a restatement of the problem of inference to the best (i.e., most plausible) explanation (i.e., model). If mechanisms for assessing plausibility could be assumed, then the problem of nonmonotonic reasoning would already have been solved, and invoking mental models would be unnecessary (for a discussion of more detailed proposals concerning how mental models might be applied to nonmonotonic reasoning, see Chater & Oaksford 1993; Garnham 1993).

What leads J-L & B to claim that they have solved the problem of nonmonotonic reasoning? Simply the observation that inferences made on the basis of a particular model may not follow deductively from the premises given; the addition of later information which demands that the set of models be searched more fully can lead to such a conclusions' being withdrawn. This certainly shows how mental models can make logically invalid inferences and later recover from them, but it says nothing about the real problem in hand. The problem of nonmonotonic reasoning is not to explain how it is possible to jump to conclusions and later revise them; this could be achieved trivially by modifying any standard method of proof. The problem is to explain how people are able to jump to *sensible* conclusions and revise these *appropriately* as new information is encountered, and indeed to elucidate what it means for a conclusion to be sensible or appropriate. Regarding these questions, artificial intelligence and cognitive science have had depressingly little to say; unfortunately, mental-model theory appears to have nothing further to add.

## Some difficulties about deduction

L. Jonathan Cohen

*The Queen's College and Sub-faculty of Philosophy, Oxford University, Oxford OX1 4AW, England*

Johnson-Laird & Byrne's (J-L & B's) arguments are beset with a number of serious difficulties. I have space to mention two.

The first can be stated quite briefly. According to the authors (p. 22), "to deduce is to maintain semantic information, to simplify, and to reach a new conclusion." But by this criterion even *modus ponens* should seem an "odd or improper" (p. 21) deduction because it does not maintain semantic information. For example, the conclusion "There is a triangle" obviously does not maintain all the semantic information contained in the premises "If there is a circle, there is a triangle" and "There is a circle." So either J-L & B are wrong to assume (e.g., p. 41) that such a *modus ponens* inference is a respectable example of deduction, or they are wrong to hold that deduction ought to maintain semantic information. If they abandon the assumption that such a *modus ponens* inference is a respectable example of deduction, they exclude what would normally be acknowledged as a veritable prototype of deduction and leave obscure what kinds of mental processes they are in fact investigating under the rubric of "deduction." But if they abandon the view that deduction ought to maintain semantic information, they must give some other explanation for most people's reluctance to treat detachment of an asserted conjunct and similar trivialities as full-blooded deductive processes.

The second difficulty is a more complex matter. Consider a proposed inference from

> If John's automobile is a Mini, John is poor, and, if John's automobile is a Rolls, John is rich

to

> Either, if John's automobile is a Mini, John is rich, or, if John's automobile is a Rolls, John is poor.

It is easy to imagine circumstances (automobile prices, customer preferences, etc.) in which the premise of this proposed inference is true and the conclusion is false. So the inference appears to be invalid, and we seem somehow to be capable of having reached this evaluation by searching in our minds for imaginable models that can function as counterexamples. But, if we were to implement J-L & B's own model-theoretic, counterexample-seeking algorithm, we ought apparently to come up with a different evaluation of the inference's validity: The inference

would apparently turn out to be valid. And this is because the authors' algorithm is apparently designed to achieve accordance with the norms of two-valued truth-functional logic, as is evident from the way in which it handles disjunctions, conjunctions, conditionals, and so on. If that is how you interpret "if," "and," and "or" in the above inference, it comes out valid. Indeed, the inference can be represented by a valid deduction in any natural deduction system for the classical propositional calculus; and the conditional formed by taking the inference's premise as antecedent and its conclusion as consequent is true under all uniform distributions of truth-values among its component propositions.

These facts have rather serious implications for J-L & B's theory of deduction. The authors (p. 19) ascribe a correct deductive competence to any normal human being. According to their view, logical mistakes in reasoning are all ascribable to accidents of performance, though such mistakes are alleged to affect some people more than others, because the complexity of the model with which people have to operate on some kinds of occasion tends to make the problem too difficult for some reasoners to handle correctly. However, if normal people are attributed a correct deductive competence, they should be supposed capable in principle of evaluating the above inference as invalid, whereas people's implementation of J-L & B's algorithm would lead to their being systematically inclined to evaluate the inference as valid.

There are perhaps three main ways in which J-L & B might seek to escape this criticism: (1) They might try to argue that the inference is in fact valid; (2) they might modify their ascription of a correct deductive competence to normal people so that normal people can then be supposed to be inclined in principle to hold the above inference valid even though it is invalid; or (3) they might try to modify their algorithm in such a way as to avoid any inclination for it to produce favourable evaluations of validity for the above inference.

(1) A claim that the inference is in fact valid would be in line with J-L & B's treatment of some of their own examples, such as the inference (p. 74) from

Shakespeare wrote the sonnets

to

If Shakespeare didn't write the sonnets then Bacon did.

But J-L & B would have to grant, in accordance with their truth-functional logic, that if this inference is valid so too is the inference from the same premise to, say,

If Shakespeare didn't write the sonnets then Essex did.

And, at least in ordinary commonsense judgement, these two counterfactual conditionals cannot both be true. Should we really be happy with a logic that licenses the deduction of incompatible conclusions from the same empirical premise?

Similarly, in the proposed inference about John's automobiles, the proposed disjunctive conclusion is a rather poor candidate for commonsense acceptance as true when the conjunctive premise is so accepted. So the trouble here is not that reasoners are unwilling to assert a proposed validly deducible conclusion because they would then be insufficiently parsimonious in their statement of the available information (p. 184). The point that needs to be emphasised is rather that the alleged conclusion is just not validly deducible at all, irrespective of whatever attitudes to semantic information play a part in the overall situation. The fallacy of thinking that the alleged conclusion is validly deducible is founded on adopting a mistakenly truth-functional interpretation of certain conditional sentences of ordinary language. Classical truth-functional logic is an elegant system of theory, but, like so many other mathematical and scientific idealisations, its application to reality is by no means a straightforward matter. The elegance and simplicity of the theory sometimes tempt people to use its concepts to construct an inaccurate description of the data, and if psychological investigators of human reasoning do not guard adequately against this temptation they are bound to fall into paralogism.

(2) Another conceivable escape route would be for J-L & B to insist that the hypothesised mental mechanism exists but is not even in principle, let alone in practice, thoroughly rational. In this view the hypothesised mechanism is capable of generating errors even without malfunction or unusual complexity. Correspondingly the authors would abandon their claim that normal human beings have a correct deductive competence. At least some types of faulty reasoning – instantiated above – would be considered to be due to the very nature of the modelling process, and it would be claimed that the mistake would regularly be made even by someone who did not find logical complexities difficult to handle.

This kind of escape route is not in fact available to J-L & B, however. The trouble is not just that they have given one or two arguments (p. 209) for supposing that deductive rationality transcends cultural differences. There is a deeper problem. The trouble is that the rationality principle is integral to much of the authors' methodology, although they do not explicitly recognise that this is so. In many cases the rationality principle underwrites the kind of interpretation they wish to impose on their experimental data.

It is important to see how this comes about. J-L & B's overall strategy is to show that the "ease" with which experimental subjects perform deduction-related tasks can be systematically explained by the hypothesis that subjects manipulate appropriate mental models, but not by the hypothesis that they operate in accordance with appropriate formal rules. J-L & B measure "ease" here mostly by the percentage of logically correct responses that are given. So they presuppose that the competence with which they are dealing – the competence that is executed by the subjects' responses – is the ability to recognise or carry out *correct* deductions. For, if the competence included a disposition to make certain erroneous kinds of deduction, the "ease" with which such deductions were made would have to be measured by the percentage of appropriately *in*correct responses given. In other words, you have to take sides on the issue of the rationality principle before you can expect to draw justifiable conclusions about the nature of the algorithm explaining deductive performance. It follows that much of J-L & B's argument would collapse if the rationality principle had to be abandoned.

(3) The other conceivable escape route for J-L & B would be to insist that they are right in ascribing a correct deductive competence to normal human beings but to admit that they are wrong about the details of the algorithm that explains human deductive performance. Somehow or other, in this view, a different type of algorithm operates, and *if* that algorithm does not involve the use of formal rules or of content-specific criteria, it must involve some nonclassical type of modelling. The inferences in question could then be judged invalid, as they ought to be. But any revised algorithm that is proposed would need careful examination if it is to justify the claim that it has explanatory cogency. The experimental data would have to be reanalysed and reinterpreted in order to show that the suitably revised model-theoretic algorithm is indeed still superior, in the cases under consideration, to a formal-rule algorithm.

There is still a long way to go, therefore, before we have an adequately supported "explanation of how logically untutored individuals make deductions," which is what Johnson-Laird & Byrne claim to present (p. x).

## Tractability considerations in deduction

James M. Crawford
*AT&T Bell Laboratories, 600 Mountain Ave., Murray Hill, NJ 07974-0636*
**Electronic mail:** *jc@research.att.com*

Johnson-Laird & Byrne (J-L & B) discuss a number of interesting examples of hard and easy types of deduction. Many classes of deductions they find to be easy for people are similar to classes of deductions that have been studied by the tractable reasoning community in Artificial Intelligence. This similarity is somewhat surprising because work in tractable reasoning has focused primarily on computational complexity rather than cognitive modeling. Certain types of deductions may simply be hard for artificial and human reasoners, and for rule-based and model-based reasoners. However, there also seem to be tractable deductions that people find difficult. It is useful to compare the classes of deductions that are known to be intractable with the classes that people find difficult. Deductions that are computationally tractable but difficult for people may provide important evidence of the types of reasoning algorithms people are using. In addition, examples of classes of deductions that are easy for people but not possible with existing tractable inference algorithms could be particularly interesting for both the cognitive psychology and tractable reasoning communities.

In general, logically sound and complete inference is undecidable. Nevertheless, some types of reasoning are computationally easy even in extremely large knowledge bases (e.g., reasoning with production rules as in an expert system). A body of recent work in tractable reasoning (Crawford & Kuipers 1991; Givan et al. 1991; Shastri & Ajjanagadde 1990) has therefore sought to divide the inference process into a tractable portion and an intractable portion. From a cognitive viewpoint, the tractable portion performs reasoning that is analogous to inferences people find obvious (and do not consider to be inferences at all), while the intractable portion performs deductions that are analogous to those requiring conscious thought. From a computational complexity viewpoint, the tractable portion comprises an algorithm that is guaranteed to terminate within a fixed time bound (usually taken to be polynomial, or in some cases [Shastri & Ajjanagadde 1990] logarithmic, in the size of the relevant portion of the knowledge base). The tractable portion is usually thought of as executing to completion on each query or assertion to the knowledge base whereas the intractable portion is thought of as being under the control of some high-level routine that reasons about the utility of performing various types of computationally expensive inference. This distinction between tractable and intractable inference is important to cognitive psychology because it suggests that one should assess the difficulty of a deduction by determining whether it requires intractable inference (ignoring, for example, the application of any number of rules from a tractable rule set – a rule set that can prove to be deductively closed within polynomial, or logarithmic, time).

There appear to be two sources of intractability in first-order inference: reasoning by cases, and reasoning with $\forall\exists$ quantification. Results in Crawford and Kuipers (1991) seem to indicate that if these two are split off then the remainder of first-order inference is tractable. It is interesting that both types of reasoning were found by J-L & B to be difficult.

Consider first reasoning by cases. This corresponds roughly to what J-L & B refer to as reasoning with multiple models. Sound and complete propositional reasoning can be done with *modus ponens* together with the ability to make assumptions and reason by *reductio ad absurdum*. Crawford and Kuipers (1991) show that the time complexity of such reasoning is exponential in the depth of assumption nesting (assumptions are nested when one is made within the context of another). This argues that the difficulty of a propositional deduction should be assessed by counting the depth of assumption nesting it requires (rather than, for example, the number of steps in a proof of it). When one makes an assumption and reasons by *reductio ad absurdum*, one splits the possible states of the world into two cases – one in which the assumption is true and one in which it is false – and deals with each case in turn. Those problems that require nested assumptions, such as double disjunction problems, are thus the ones that require multiple models. Such problems were found by J-L & B to be among the hardest propositional reasoning problems.

The undecidability of full first-order logic can be traced back to the existence of $\forall\exists$ quantification (i.e., existential quantifiers within the scope of universal quantifiers). Such quantification may force models of logical theories to contain an unbounded number of individuals (more precisely, without such quantification one can place an a priori bound on the size of models that need to be considered). To see intuitively why this is so, consider the assertion "every man has a father who is a man." Such an assertion allows the creation of an unbounded string of fathers of fathers. Crawford and Kuipers (1991) suggest that $\forall\exists$ quantification should be dealt with by preventing the tractable portion of the inference algorithm from creating new conceptual objects. J-L & B also find problems involving $\forall\exists$ quantification to be appreciably harder in general than problems involving other kinds of mixed quantification. They claim, incorrectly, that there is no intrinsic logical difficulty in $\forall\exists$ quantification. To support this claim they give an example of a simple transitive deduction from two $\forall\exists$ premises that they show is relatively easy for people to make. However, this deduction is so easy (and so easy to get right by a simple-minded application of transitivity that ignores the quantification completely) that it does not address the deeper difficulties inherent in reasoning with $\forall\exists$ quantification.

Tractable inference algorithms were not intended to be cognitive models, and there are cases in which the performance of existing tractable inference methods is clearly superhuman. For example, it is not clear that *modus tolens* is any less tractable than *modus ponens*, but people seem to find deductions based on *modus ponens* easier. In addition, the tractable rule set in Givan et al. (1991) supports syllogistic reasoning that J-L & B have found quite difficult for people. This brings up the interesting question of why human commonsense reasoning has not come to use the strongest possible tractable inference methods. There are (at least) three possible answers to this question: First, human reasoning may be constrained in ways the preclude the use of existing tractable inference methods (e.g., by limitations in working memory size); second, it may be that in the "real world" commonsense reasoning would not significantly benefit from stronger inference techniques; and, finally, human reasoning may simply be optimized for other more important types of inference.

In general, J-L & B could do a better job of distinguishing between the types of reasoning that are computationally complex, and those that are difficult because of the particular reasoning algorithms people use. Their analysis of rule-based inference systems would also be more convincing if instead of merely counting the lengths of proofs, they distinguished between the application of rules from tractable rule sets (rule sets that can be deductively closed in polynomial time) and the application of intractable proof rules (see Givan et al. 1991). Nevertheless, neither of these points invalidates J-L & B's basic argument that people reason by building models. The book is well worth reading, both for the wealth of experimental studies it reports on and for the insight the authors provide into human commonsense reasoning.

# Deduction by children and animals: Does it follow the Johnson-Laird & Byrne model?

Hank Davis

*Department of Psychology, University of Guelph, Ontario, Canada N1G 2W1*

**Electronic mail:** *psyhank@vm.uoguelph.ca*

George Bernard Shaw (1933) based his indictment of Pavlov and, by implication, of experimental psychologists in general, on the fact that they studied things that were obvious to any schoolchild. Why invest time, money, and suffering, he argued, in unravelling processes that were already clear?

A good example of behavior that might seem to require little explanation is how humans use logic. We all know that humans frequently behave in a reasonable or rational manner and we have known the rules of formal logic since Aristotle. Why not simply conclude, as Inhelder and Piaget (1958) did, that "reasoning is nothing more than the propositional calculus itself"?

After several years of progressively chipping away at the foundation of this belief, Johnson-Laird & Byrne (J-L & B) have written a book that finally demolishes it. The demolition is handled deftly. Not only has the structure of the old theory been razed, but a new one has been erected in its place.

This is a landmark book, not necessarily because it is correct, but because it represents the first attempt to develop a comprehensive theory that explains all facets of deduction. What makes J-L & B's theory noteworthy is that it is formulated without reference to the rules of inference. The authors will be roundly criticized both for demolishing the old view and for the specifics of what they have created to replace it.

Arguably, J-L & B have fashioned a less inspiring view of human cognition. There is an undeniable elegance to the rules of formal logic and, although humans are inherently capable of mastering the propositional calculus, according to J-L & B, they have not done so. They have instead constructed models or representations of the facts and gone looking for disconfirming examples. They have substituted a rough form of hypothesis testing for abstract logic. In this view, humans are, in a sense, "beating the system." They are not unlike Clever Hans, making a devious adaptation to the demands of his world, performing well enough, yet mastering nothing about the basics of mathematics. This is not a view of deduction with which to aggrandize the intellect of our species. But it is probably correct.

Having said this, I will direct my attention to a less central aspect of J-L & B's work: its relevance to the deductive behavior of nonadult humans. From the point of view of comparative psychology, the timing of this book could not have been better. During the past decade, a new literature has emerged that shows nonadult human subjects to be capable of deductive reasoning. These experiments focus on transitive inference (TI), a form of deduction previously reserved for adults, or at least for those at an advanced stage of cognitive development (Piaget 1970).

Studies of transitive inference typically consist of paired comparisons such as *Alan is taller than Bill, Bill is taller than Charles,* which lead to the inference that Alan is taller than Charles, although the two have never been directly compared. Ground-breaking experiments with children (e.g., Bryant & Trabasso 1971; Riley & Trabasso 1974) have demonstrated success when procedural accommodations were made for these young subjects. This work suggested that what were presumed to be cognitive deficits may actually have been memory problems. Such evidence of "adult" cognition in children prompted a related question: Why not animals?

Assuming that verbal abilities were not critical to such tests, Gillan (1981) devised a procedure for chimpanzees in which paired comparisons used food reward to indicate which of two colors was the "winner." Thus, for example, *red is rewarded over blue* becomes the analogue to *Alan is taller than Bill.* In this manner, a five-element series of colors was established using a sequence of paired comparisons (e.g., red > blue; blue > yellow; yellow > green, etc.). This training was presumed to yield an ordered mental representation ($A>B>C>D>E$) which allowed the subject to draw inferences when tested on the novel, nonadjacent $B$ versus $D$ comparison (correct response = $B$). Gillan's chimpanzees succeeded on this task, as did McGonigle and Chalmers's squirrel monkeys (1977), von Fersen et al.'s pigeons (1991), and rats in my own laboratory. The latter were tested using olfactory rather than visual stimuli, but succeeded nonetheless (Davis 1992).

Once we concede that children, chimpanzees, and rats are capable of making transitive inferences, it is clearly time to question the basis on which such deductions normally occur in adults. Putting the case simply, if a mechanism common to all species is involved, either adults must be barely engaging their cognitive abilities or rats or pigeons must be stretching theirs to the breaking point. It is here that the value of J-L & B's book to comparative psychology becomes most apparent. Ironically, their book was not written in an attempt to account for the performance of "lower" subjects. Nevertheless, if human deductions can be demystified in a way that eschews the rules of predicate calculus, then accounting for TI in nonadult humans may not be insurmountable.

The question that will face comparative psychologists is whether the alternative, nonsyllogistic model of deduction proposed by J-L & B need be common to all species. Should we accept the basics of their view and search for ways in which animals might form representations of their environment and look for disconfirming evidence? J-L & B's book might conceivably have an altogether different effect on comparative cognition. Now that we have been freed from the yoke of classic Aristotelian dogma, can we not assume that deduction can be achieved in a variety of ways and carry out our search without preconception? Given the history of comparative psychology, this scenario seems less likely.

To date, attempts to "explain away" TI in animals, who were thought incapable of formal logic, have taken two forms. The first is based solely on associative learning and makes no reference to mental states or deductive capabilities. This behaviorist position assumes that successful performance can be understood in terms of Pavlovian processes inherent in the way premise information is trained. Correct selections, thus, reflect not inferences, but movement toward stimuli of greater associative strength or value (e.g., von Fersen et al. 1991). The second alternative account of TI may be described as a form of spatial paralogic, in which representations of premise information are thought to be mapped in spatial form. Thus, tests requiring the subject to evaluate novel premises can be viewed as perceptual rather than cognitive tasks (e.g., Huttenlocher 1968).

It is difficult to dismiss the latter account of deductive performance in animals for precisely the same reason that J-L & B's model may prove hard to evaluate in nonadult humans. How will the tenets of this theory be implemented in nonverbal terms? As previous TI testing of animals indicates, the task is not insurmountable, but it will require more than a little cleverness. It also risks stirring up unwarranted questions about whether animals "think" (cf. Davis, in press; Griffin 1974).

If their book is successful and the J-L & B model becomes the dominant account of deductive competence in humans, its application to nonadult subjects will be almost inevitable. This is how comparative psychology operates. As a researcher whose primary concern is with animal cognition, I can only hope that the model's evaluation with nonverbal subjects will be as rigorous as the work that went into its development with adults.

## Mental-model theory and rationality

Pascal Engel
*CREA, Ecole Polytechnique, 75005 Paris, France*
**Electronic mail:** *engel@poly.polytechnique.fr*

Johnson-Laird & Byrne (J-L & B) argue that mental-model theory (henceforth MMT) allows us to "resolve" the controversy between rationalists and relativists about the nature of rationality, claiming that there is a "central core" of rationality, namely, the principle of semantic validity (pp. 207–9). I find, however, that their treatment of this topic is too elusive and unspecific to be truly convincing. I happen basically to agree with them, but I would like to suggest a better way of defending an idea similar to theirs.

In the first place, it would be good to distinguish at least two concepts of rationality (in the manner of Evans, in press). The first is rationality of purpose (rationality$_1$): adaptation of means to ends in order to achieve the best result (e.g., maximising expected utility). The second concept is rationality of process (rationality$_2$): What are the requisite features of the psychological processes people use to be rational? In general, rationality$_2$ is assimilated to logicality and especially to deductive logical competence. Although they agree that thinking in general is not confined to logical competence (p. 206), J-L & B have a tendency to reduce the issue of rationality versus relativism to the issue of deductive competence, and therefore to rationality$_2$. MMT, however, allows us to cope with rationality$_1$: This is best achieved is we suppose that people use models in their thinking, rather than specific rules of reasoning. In this respect, the Craikian (1943) claim that human thinking is the manipulation of models could be interpreted in light of the view that the evolution of human thought is supervenient upon the development of a mimetic culture through which humans have learned to use representations and models of the world (Donald 1991). I concur with J-L & B's extreme caution with such claims, however, especially when they are associated with the view that specific logical rules could be accounted for by the use of evolved "social contracts" (Cosmides 1989; J-L & B 1991, p. 78).

J-L & B seem to agree with Barnes and Bloor's (1982) criticism of the notion of universal criteria of deductive rationality, which is based on Lewis Carroll's story of Achilles and the tortoise. The story is supposed to show that there can be no noncircular universal justification of deduction, because any attempt to justify a single logical rule (e.g., *modus ponens*) presupposes the validity of the very rule. But the circle can also be taken to be a nonvicious one if one adopts Goodman's (1966) or Rawls's (1970) well-known notion of a "reflective equilibrium." According to such a view, we posit various normative principles; then we revise them when they conflict with our intuitions until we reach an equilibrium between principles and intuitions. Cohen (1981) uses this strategy to show that logical competence cannot but agree with universal standards. In this view, the very evaluation of a logical system presupposes that we use the standards of deduction we are going to ascribe or deny to people according to whether they deviate from those standards. There is hence a core of logical competence that belongs to our very practice of interpreting logical or illogical beliefs (a similar idea is advocated by those writers who propose that we interpret others through a general "principle of charity").

Johnson Laird & Byrne identify the core of logical rationality with some sort of cognitive universal, the semantical principle of validity: An argument is valid only if its premises cannot be true while its conclusion is false. This accounts for the search for counterexamples that, according to MMT, is the basis of our uses of mental models as well as the possibility of error or of deviancy from the standards of normative logic. In the view suggested here, the common core of rationality operates both as a cognitive universal, true descriptively of individuals, and as a normative principle for the evaluation of logical reasoning in experimental research. The fact that the semantic principle of validity is a cognitive universal allows us to answer a criticism that has been levelled against the reflective equilibrium method: Truth could be only one among our multiple interests in our evaluation of cognitive tasks (Stich 1990), another interest being utility. But if logical reasoning is, as a matter of psychological fact, an evaluation of truth among models, it is both pragmatically useful and cognitively adequate to search for truth.

## On rules, models and understanding

Jonathan St. B. T. Evans
*Department of Psychology, University of Plymouth, Plymouth PL4 8AA, England*
**Electronic mail:** *po2118@pa.plym.ac.uk*

Johnson-Laird & Byrne (J-L & B) seek to distinguish their own account of deduction from two versions of rule-based reasoning: the use of formal rules in the form of "mental logics" and the induction of content-specific rules and schemas. The mental-logic theory is criticised for – among other things – its failure to explain content effects in reasoning, whereas the pragmatic schemas approach is conversely criticised for failing to account for deductive competence with abstract material. The authors, however, seem on the whole to be less certain in their opposition to the content-dependent rules. For example: "Knowledge undoubtedly influences deduction, but is it represented by content-specific rules? There is no evidence for this form of representation; it could be represented by general assertions, which are used to construct models" (p. 79).

As you read through the many and varied experimental demonstrations of mental-model reasoning offered by J-L & B, it is quite striking how (with one exception: Oakhill et al.'s work on belief bias) the studies have used arbitrary problem content. Thus, although the mental-logic theory has been criticised for its failure to account for pragmatic influences in deductive reasoning (see Evans 1991), the great majority of the mental modellers' own experiments seem intentionally designed to exclude these same influences. This does seem to be the area of weakest development of the theory. For example, the massively researched effect of content on the Wason selection task is given perfunctory treatment (pp. 77–81), with little elaboration of the proposed mental-model explanation.

In exploring this area of uncertainty I would like to make use of two distinctions: one concerning reasoning with novel and familiar problems and one concerning implicit and explicit cognitive processes. My contention is that the mental-model theory provides the most plausible account of how we reason with novel material, such as that presented in the great bulk of the experiments run by Johnson-Laird and his colleagues. By definition, we cannot apply rules or schemas induced from previous experience. I find the mental-logic account implausible for much the same reasons as do J-L & B. So the subjects must indeed laboriously consider all the possible states of the world in which the premises could be true, and try to find a novel conclusion that applies in all of them. This (mental-model) account is a description of what I regard as reasoning from first principles.

At the same time, I find the mental-model theory a most implausible account of how we reason with familiar material. The exhaustive construction of models necessitated by the search for counterexamples requires high cognitive effort and induces many errors when only two or three different models need to be constructed, as in the case of syllogistic reasoning (Johnson-Laird & Bara 1984). Reasoning with semantically rich material could involve many more models – leading to computational intractability. However, the key point is that in reasoning with familiar material it is much more efficient to utilise domain-

specific rules and schemas induced from previous experience.

The distinction between implicit and explicit cognition is relevant here also. I do not believe that we utilise explicit (conscious, reportable) rules for reasoning any more than we do for language comprehension or visual recognition. Retrieval and application of schemas, for example, is a recognition type of process that might well be modelled by a PDP network (see Rumelhart et al. 1986). We know that people can acquire, by experience, effective rules for controlling actions without explicit, verbal knowledge of the rules (e.g., Berry & Broadbent 1984) and that, conversely, possession of explicit rules can be ineffective in actions (Broadbent et al. 1986).

Let us consider the example of a student learning procedures for statistical significance testing. Some students will demand to be given explicit rules describing when one test should be used rather than another. This approach is rarely successful, leading to a blind "cookbook" approach. The rules are either too simple, so that they do not apply in all cases, or they are too complex to remember. In either event, they are applied without understanding. So what do good statistics teachers do? They encourage the student to think through the range of possibilities, to consider the possible forms the data set might take, the various statistical tests available, the restrictions that apply to each usage, and so on. Eventually, by a process of elimination, the student can work out (deductively) the correct procedure.

The process described above is a first-principles, mental-modelling type of procedure. Experts can work at this level – in order to teach – or in order to deal with complex problems of a novel kind. However, they would most certainly not reason in this way when making routine decisions about statistical tests applied to research problems of familiar kinds. Such decisions are normally made rapidly and "intuitively." This is because content-specific rules or schemas induced from past experience are being applied, and because such cognitive processes are implicit in nature.

Curiously enough, it is this intuitive grasp of concepts that we refer to by the term "understanding." Understanding does not imply the ability to explicate – as the knowledge engineers learned to their cost in the quest for expert systems. The use of explicit rules is the very opposite of understanding – it is cookbook thought. One is reminded here of Wertheimer's (1961) distinction between blind and productive thinking.

So, in conclusion, the theory of deduction by mental models provides a very promising and important insight into our ability to reason with novel content. The ability – when required – to reason by explicit modelling of possible states of the world is indeed a vital facet of human intelligence. It allows us to attempt novel problems and helps us to acquire understanding. It must be recognised, however, that this is probably not the mechanism by which most of our inferences – relating to familiar material – are made.

## On modes of explanation

Rachel Joffe Falmagne

*Department of Psychology, Clark University, Worcester, MA 01610*
**Electronic mail:** *rfalmagne@vax.clarku.edu*

Reality is complex. Especially so are cognitive processes, including those underlying deduction. What then is a useful approach to psychological theory? I see two guidelines, one substantive and one metatheoretical. From a substantive standpoint, at this particular juncture the pressing need is for an integrated analysis of deduction, one that articulates the interplay of different kinds of processes and representations in the deductive process. From a methodological or metatheoretical standpoint, the road to a mature understanding of cognition lies in the integration of complementary theoretical frameworks, not in controversies between radical views.

Despite the many merits of Johnson-Laird & Byrne's (J-L & B's) book, the approach it takes is questionable on both counts. Although the theoretical constructs are valuable and the empirical findings informative, the discussion is stuck in the "right/wrong" rhetoric so characteristic of the current discourse of cognitive science. The strategy is to radicalize the analysis of a psychological phenomenon into extreme, mutually exclusive opponent accounts, each claiming to be capable of explaining the phenomenon in its entirety with a single theoretical language and one or two constructs. One then devotes the bulk of one's analytical and rhetorical energy to establishing the supremacy of one extreme account over the other. Thus, J-L & B construe the issue as a competition between theories relying (exclusively) on mental models or "rule" theories; their rhetoric is one of advocacy. Is this really the most fruitful strategy in the long run? I do not think so.

Any "radical" theory attempting to account for a complex cognitive process in terms of a single theoretical language has built-in limitations in explanatory range. The "advocacy" rhetoric is misguided in presupposing that the adequate explanation resides on one side or the other, when the actual relation between theories is often one of complementarity. The contrast between symbolic theories and connectionist models offers a prime example. Symbolic theories are well-equipped to describe the central part of the cognitive process computationally but ill-equipped to deal with the interface between the cognitive process and the world. Conversely, connectionist models are well-designed for capturing the input and output end of the cognitive process but it is unclear at this point how capable they are of building complex cognitive structure. These models have typically been thought of as rivals and the rhetoric has been one of competition. Within an integrative perspective, however, a more fruitful move would be to let each model do the job it is best equipped to do and to develop the interface between the two theoretically. (See Falmagne, 1992, for an elaboration of this point with reference to issues of acquisition.) Likewise, the interest of the mental-model construct notwithstanding, it seems likely that an adult, and perhaps a seven-year-old, upon hearing "If Mary thabbles, then she fibbles" and "Mary thabbles," will conclude that Mary fibbles, a deduction undoubtedly relying on the form of the argument. It is simply not possible to perform the speedy elaborate inferences that underlie much of our mental life without some kind of automaticity and reliance on form. Given that we know the syntax of our language, and given the extensive analogies between syntax and logic as alternative formalizations of natural language, it seems very likely that we also have some formal representation of certain deductive principles (Falmagne 1988). To deny this is to deny the human mind its capacity for abstract thought.

Thus, what is needed is an integrated account of deduction that articulates how semantic and formal (or, loosely speaking, syntactic) processes interact in deduction and that articulates the interplay between logical and extralogical ingredients of reasoning. The mental-model construct is interesting and well documented, but it is bound to capture only part of the story. J-L & B's book does a fine job in consolidating and elaborating Johnson-Laird's prior proposals across a range of cognitive domains and in spelling out the theory more explicitly than was the case for certain situations, but both the theory and the evidence are more convincing for, say, the highly elaborate reasoning involved in multiply quantified sentences than for simple instances of propositional reasoning.

Finally, the cognitive status of mental models within the overall mental organization requires comment as well. I have no doubt that something like mental models is involved in deductive processes at some level. But in accounting for deductive activity, a fundamental distinction must be drawn between two levels of mental representation, a deep level of knowledge representation where logical knowledge "lives," so to speak (whether as formal principles, semantic representations of con-

nectives, or procedures), and a functional, on-line level of representation, where the action happens during actual deductions. When an inference has to be carried out, a functional representation of the problem must be constructed in working memory. To speculate, this functional representation may either highlight the form of the argument or it may consist of a semantic model or of an image, depending on a number of factors, but that is not the main point here. The point is that the mental models described by Johnson-Laird & Byrne are functional representations constructed on-line and one must therefore clarify the procedures underlying their construction. The question is particularly problematic for multipremise inferences, because the interfacing of the individual models with one another, in order to be done properly, must be based on an overall understanding of the constraints of the problem, going beyond the semantic representation of each premise and must be monitored by a logical executive function.

## The argument for mental models is unsound

James H. Fetzer

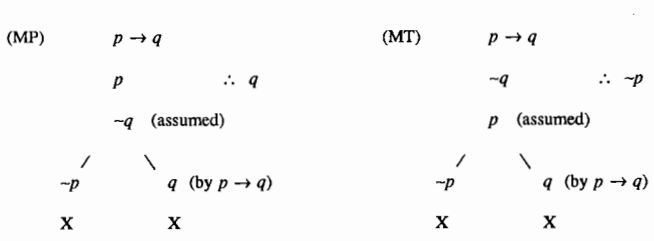
*Department of Philosophy, University of Minnesota, Duluth, MN 55812*
**Electronic mail:** *jfetzer@ub.d.umn.edu*

The argument advanced by Johnson-Laird & Byrne (J-L & B) (which is summarized in Ch. 10) assumes the following form. Human performance with respect to deduction involves either formal rules, content-specific rules, or mental models. But it does not involve formal rules and it does not involve content-specific rules. Therefore, it involves mental models. Formulated in this fashion, the argument appears to be valid, since their conclusion would have to be true if their premises were true. However, there seem to be at least three reasons why their premises should not be accepted at face value.

***1. They only consider systems of natural deduction.*** The first reason is that J-L & B's rejection of formal rules is restricted to the conception of these rules in systems of natural deduction. As Blumberg (1967) has observed, there are at least three kinds of systems of formal rules, among which natural deduction systems are only one. Axiomatic systems and tree systems are alternative kinds of systems of formal rules, which the authors do not explicitly consider. Even if their argument for the elimination of models of natural deduction were a complete success, they would not have established that human reasoning does not rely upon formal rules.

Although tree systems are sometimes viewed as special cases of systems of natural deduction, their applicability has an intuitive appeal which systems of natural deduction do not share. This method requires making a list of the premises of an argument and of the negation of its conclusion in a standardized form. The truth conditions for each premise are then added to the list, where the disjunction of $p$ and $q$, for example, is represented by a branch. An argument turns out to be valid just in case every branch that is generated by repeating this process contains a sentence and its negation.

Consider, for example, arguments of the form *modus ponens* in relation to those of the form *modus tollens*. Their trees would look like this, where "X" means that the branch thereby generated is "closed" (or inconsistent):




(A nice introduction is found in Gustason & Ulrich 1973, Appendix B.)

It might be thought that overlooking tree systems, for example, makes no difference to J-L & B's position. Yet they maintain (on p. 195) that the difficulty of a deduction is supposed to depend upon "two factors: the number of steps in the derivation, and the relative availability, or ease of use, of the rules used in the derivation." They support this claim with examples intended to show that *modus tollens* is "harder" than *modus ponens* because it requires "a longer derivation." These sample tree proofs, however, provide a clear indication of the untenability of their contention. (Indeed, their own illustration improperly compares a direct proof of a *modus ponens* argument with an indirect proof of a *modus tollens* argument!)

More important, J-L & B's discussion does not appear to appreciate adequately that formal systems are models too. Formal inference rules are intended as suitable representations of underlying semantical relations. Principles codified by formal calculi (such as the predicate calculus) are intended as syntactical models that provide a valuable method for evaluating an instance of reasoning, provided that reasoning has been cast into a standard (or "canonical") form on the basis of its syntactical and semantical properties. These principles are not supposed to be directions for thinking.

The reason that these syntactical principles of inference may be useful for this role is that they have been deliberately constructed to reflect relevant semantic considerations without which they would be theoretically insignificant. The rule *modus ponens*, for example, applies only to substitution instances of $p$ and $q$ that are declarative sentences that are either true or false. It does not apply to imperatives, interrogatives, or exclamations. Its application presupposes that the same words that appear in the premises must have the same meaning as they have in their conclusions. Its very acceptability hinges upon its invulnerability to counterexamples.

Beth (1955/1969), for example, distinguishes between semantic entailment and syntactical derivability, where semantic entailments obtain between premises of the form $p$ and conclusions of the form $q$ whenever the falsehood of $q$ is incompatible with the truth of $p$, whereas syntactic derivability obtains whenever $q$ follows from $p$ in accordance with accepted syntactic rules. The purpose of *soundness proofs* for formal systems thus becomes that of establishing that every permitted relation of syntactic derivability is justified as an instance of semantic entailment. And *completeness proofs* are aimed at insuring that every permissible semantic entailment is allowed by corresponding syntactical rules insuring that it is syntactically derivable. Formal rules, like mental models, are justified by means of semantics.

***2. Their argument could just as easily be reversed.*** The second reason is that J-L & B's rejection of formal rules and of content-specific rules appears to be based upon an indefensible conception of their function in relation to human performance. J-L & B suggest that content effects sometimes affect the conclusions people draw, which allegedly undermines the development of procedures for translation into the logical forms required by formal rules, on the one hand, and on the other, that people sometimes draw inferences on the basis of considerations of logical form, which allegedly undermines their reliance upon content-specific rules.

Notice how both of these arguments could just as well be reversed in support of the opposite conclusion. Because people sometimes draw inferences on the basis of considerations of logical form, human performance sometimes (1) satisfies the logical forms required by formal rules; and (2) because content effects sometimes affect the conclusions that people draw, the performance they display sometimes satisfies the content-specific rule conception. There appears to be no more justification for interpreting this evidence in support of one conclusion rather than the other.

J-L & B's rejection of formal rules and of content-specific rules, moreover, is predicated on the assumption that the proper purpose of these rules is to account for "all aspects of performance" (p. 35), which appears to be more than could be asked of any system of rules or of procedures, no matter how adequate. After all, performance can be affected by motives, beliefs, ethics, abilities, and capabilities that go far beyond appropriate rules of reasoning. Heart attacks (phone calls, etc.) might affect "aspects of performance," but surely they need not be encompassed by an adequate theory of deduction.

***3. Their position hinges upon a crucial equivocation.*** The third reason is that J-L & B's enthusiasm for mental models appears to hinge upon a crucial equivocation. The "final step" in the construction of mental models is supposed to be to search for alternative models that refute them, where "The conclusion is valid if there are no such counterexamples" (p. 196). Observe, however, that the existence of counterexamples is perfectly compatible with the failure to discover them. Persons engaged in the construction of mental models might actually believe that their arguments are valid when they are invalid, or that they are invalid when they are valid.

Human beings, after all, display various degrees of logical acumen. The existence of a counterexample to a specific mental model may or may not be noticed by a certain thinker on a certain occasion. That is obviously an important – indeed, crucial – "aspect of performance." Unless J-L & B are prepared to deny the difference between *merely believing that an argument is valid* and *that argument's being valid*, they ought to admit that the theory of mental models supplies an account of when an argument is thought to be valid rather than of when an argument is valid.

A theory of validity for mental models should no doubt distinguish kinds of models on the basis of their form and provide general patterns that distinguish valid from invalid models. This requires a theory that provides (i) procedures for translating mental models into standardized form; (ii) indicators for determining when a model is intended to be deductive; and (iii) standards that apply to evaluate the validity of mental models on the basis of their form. In all of these respects, the theory of valid mental models qualifies as a system of rules.

In defense of their position J-L & B will of course want to insist that their theory is intended to be psychological and descriptive of human performance, whereas formal systems are instead intended to be philosophical and normative of human competence. But that defense, I am afraid, "proves too much," because it affords a foundation for understanding that the function of systems of logic is to codify the conditions of validity for arguments in the same sense in which a counterpart theory would codify the conditions of validity for mental models.

## Deductive reasoning: What are taken to be the premises and how are they interpreted?

Samuel Fillenbaum

*Department of Psychology, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599-3270*
**Electronic mail:** *usamfi@unc.bitnet*

"Some inferences are so easy to make that their validity is indisputable" (Byrne 1989, p. 62). Given "If it is raining she will get wet" and told "It is raining," subjects overwhelmingly (96% of the time) conclude "She will get wet." But if an additional premise "If she goes out for a walk then she will get wet" is added to the initial premise, then subjects who are told only "It is raining" are far less likely to conclude "She will get wet" – this inference is now made only 38% of the time. In a nice series of experiments Byrne (1989) has shown that *modus ponens* can be blocked under some circumstances.

What is the significance of such a finding? Johnson-Laird & Byrne (J-L & B) argue that such an "effect of content challenges the foundation of all formal theories: the assumption that the rule of modus ponens is part of mental logic" (*Deduction*, p. 81) and that it shows "that people do not have a secure intuition that modus ponens applies equally to any content" (p. 84) and therefore supports their thesis "that people make deductions not by following such rules but by building [mental] models" (p. 84). Politzer and Braine (1991) contest this claim, arguing that the problems used by Byrne are such that their content and world knowledge may lead to doubts regarding the truth of one of the premises, due to its inconsistency with other premises, and that "responses to inconsistent premisses cannot count as suppression of valid inferences," to quote the title of their paper. In turn, Byrne (1991), replying to Politzer and Braine, shows by a close examination of the particulars of their account that it makes incorrect predictions for her data. Where does all this leave us?

It may be that Byrne's work is of interest and importance not so much because it provides strong evidence against the mental-logic approach and for a mental-model approach, but rather in highlighting some issues in deductive reasoning common to both approaches, and perhaps required by any approach. J-L & B argue that any account suggesting that reasoning depends only on formal rules of inference that operate in a syntactic way on the abstract logical form of the premises is in principle insufficient because "there must be an initial comprehension step in which the premises are decoded into the representations used by the rules" (Byrne 1989, p. 64). Byrne grants that if arguments containing additional antecedents are represented in a conjunctive way so that premises of the form "If $p$ then $q$," "If $r$ then $q$" are represented as "If $p$ and $r$ then $q$," then in a mental-logic account *modus ponens* would be blocked unless both $p$ and $r$ were asserted. Thus, her results could be handled by such an account, but this requires that arguments of the same form ("If $p$ then $q$," "If $r$ then $q$") be represented differently depending on their semantic content, that is, whether the reasoner understands $p$ and $r$ as additional or alternative antecedents. So "the burden of explanation shifts from the formal rules to the process of comprehension [of the premises]" (Byrne 1989, p. 77). And J-L & B claim that the mental-logic approach has very little useful to say on such matters: "Despite almost three decades devoted to their development no such mechanism has been provided" (Byrne 1991, p. 77).

If we look at the Politzer and Braine (1991) defense of mental logic it is clear that their analysis, claiming that the problems Byrne used led to doubts as to the truth of a critical premise because of its inconsistency with other premises, must go beyond any purely syntactic, formal approach and must appeal to semantics and all sorts of matters of world knowledge. Actually, they are very explicit in agreeing with Byrne "that theories that invoke a mental logic need to be supplemented with a detailed account of the interpretive process" (Politzer & Braine 1991, p. 107). Indeed they in turn "insist that the mental-models theory lacks a detailed account of the interpretive process . . . offers no explanation of how people establish an appropriate model" (1991, p. 107). It is very likely that both the mental modelers and the mental logicians are completely right on this basic issue, namely, that the interpretive component is a critical one in reasoning and that neither position has enough revealing things to say about it. Given a sufficient account of the interpretive processes, either a mental-logic or a mental-models approach may work; without such an account, neither can.

I want to make only one more point, about the ecology of (deductive) reasoning studies and how what goes on in the laboratory may be different in some important ways from what goes on in the world outside. In the laboratory a subject will

generally assume that the experimenter is being cooperative and provides all and only the information that needs to be considered in doing the task at hand (perhaps basing these assumptions on some proto-Gricean conversational principles of quantity and relevance). Given only some premises, we are to confine ourselves to just these, given some additional premises these too must be taken into consideration, *but nothing else*. Outside the laboratory the reasoner often may not know what are (to be) the relevant considerations (and there is no such thing as "the reasoner" but rather different reasoners with different assumptions and knowledges). So reasoners in the laboratory may come to different conclusions from reasoners outside, not so much because of differences in the machinery of reasoning as in the contents that machinery operates on. Reminding us of this may perhaps be the most important contribution of Byrne's work. As she puts it in the conclusion of her 1991 paper, "in daily mental life there are always background conditions necessary for an outcome that can be called into question; and there are always alternative conditions sufficient for an outcome that can be introduced. Conditionals are elliptical" (Byrne 1991, p. 77). How the ellipses are filled will critically affect the conclusions that are reached.

## Mental models and informal logic

Alec Fisher

*Center for Critical Thinking and Moral Critique, Sonoma State University, Rohnert Park, CA 94928*

**Electronic mail:** *alec.fisher@sonoma.edu*

In many ways Johnson-Laird's work on mental models is so persuasive, and has become so widely accepted, that it is easy to take it for granted and to forget the orthodoxy it undermines. In criticising mental models, I do not wish to say anything in favour of the old view. My complaint is rather that the mental-model tradition does not go far enough in rejecting old ideas: Given its ambition to provide a general account of the way people reason, it needs to recognise that most reasoning which people actually engage in is utterly unlike syllogistic reasoning or simple propositional logic reasoning, and is not deductive reasoning at all, nor even inductive reasoning. What the mental-model tradition needs to do now is to abandon the old, logician's paradigm and look instead at the kinds of reasoning being studied by those working in the developing fields of informal logic, critical thinking, and argumentation theory. First, however, a few words about deduction.

Formal logic books usually give a few examples of deductive reasoning in the course of explaining the theory of deductive logic. The striking thing about most of these examples is that no one would ever dream of using such an argument in any real argumentative situation. Let us use the term "real reasoning" to refer to arguments which are to have been used with a view to convincing others – reasoning which occurs in everyday situations, in letters to the editor, in newspaper articles of the "analysis" variety, and in scholarly journals. A quick survey of such "real reasoning" will show that it is almost never deductive.

Here are two examples from letters to the editor:

> Our 40,000 GI's stationed in North Korea support a corrupt regime. The savings in dollars which would result from bringing them home could make a sizable dent in the federal deficit. Furthermore, the Korean conflict ended over thirty years ago. That is why it is time we brought the troops home. (*Sunday Star-Ledger*, Newark, NJ, USA)

> In Great Britain more than 200,000 people die each year from heart disease, whereas about 40,000 die from lung cancer. Heavy smoking approximately doubles one's chances of dying from heart disease and increases the chance of dying from lung cancer by roughly ten times. Most people will conclude that a smoker is more likely to suffer from lung cancer than to suffer from heart disease. Indeed, both in Britain and elsewhere, government campaigns against smoking have been largely based on this assumption. But it is clearly false. (*The Guardian*, London, England)

In recent years many philosophers and logicians, realizing that traditional formal logic has almost nothing to say about analyzing and evaluating most real arguments, have been studying such argumentation – in what has become known as the Informal Logic and Critical Thinking (ILACT) movement in North America and the Argumentation tradition in Europe (for further information, see the journal *Informal Logic*, published by the Philosophy Department, Windsor University, Windsor, Ontario, Canada). There are two things especially worth saying about the ILACT tradition in this context: (1) Almost no one working in the ILACT movement knows anything about the mental-model tradition, or pretends to know what psychological mechanisms underlie real reasoning. More important to most readers of this review: (2) Very few people working in the mental-model tradition appear to know what is happening in the ILACT movement. For example, at a recent cognitive psychology conference on reasoning in Aberdeen, Scotland, all but one of the papers were about syllogistic or simple propositional logic reasoning.

Since syllogistic reasoning and simple propositional logic reasoning almost never occur in real argumentative contexts, it seems clear that cognitive psychologists who want to study human reasoning should study what I have called "real reasoning." Real reasoning is often incomplete, often assumes things, can often be strengthened, (or weakened). Here is an example:

> In a recent study it was found that left-handed car drivers were 85% more likely than right-handers to have an accident whilst driving; left-handers were 50% more likely to have accidents in the home or whilst using tools; and they were 20% more likely to experience work-related accidents or suffer a sporting injury. Since 13% of 20 year-olds are left-handed whilst only 1% of eighty year-olds are, the conclusion is clear, as is the explanation . . . [the investigators explained the phenomenon by reference to the "fact" that the world is arranged for the convenience of right-handers]. (*The New Scientist*, London, England)

Real reasoning often appeals to principles of a general kind, for example, principles relating to the credibility of evidence ("He's a more reliable witness because . . . "). It also often appeals to the idea that arguments are parallel ("You might as well say that . . . "). It often contains considerations of counter-indications to the main thrust of the argument, as in the following example (which is couched in dialogue form so that the counterindication is easily identified, but it would be simple to paraphrase it in plain prose);

> A. Peter is clutching his stomach, he's groaning terribly, and there's blood on his hands. He must be badly hurt.
> B. Unless it's ketchup and he's acting.
> C. No, it's real blood, OK, and anyway, he's a hopeless actor.

So far as I am aware, cognitive psychologists have not studied empirically the kind of reasoning that is involved in cases such as those just described. It may be that mental models can cope well with real reasoning, with its incompleteness, with its assumptions, with what would strengthen it, with what principles are being used, with the idea of parallel structure that is appealed to, with counterindications, and so forth. Indeed, there is good evidence that it can cope with at least some of these. But the time is ripe for the mental-model tradition to turn its attention to the examples of reasoning that are the stock in trade of the ILACT tradition, namely, real reasoning, and to show that it applies here. Examples of good source books in that tradition are Fisher 1989; Freeman 1987; Govier 1981; Thomas 1986.

## Why study deduction?

Kathleen M. Galotti[a] and Lloyd K. Komatsu[b]
*Department of Psychology, Carleton College, Northfield, MN 55057*
**Electronic mail:** [a]*kgalotti@carleton.edu*; [b]*lkomatsu@carleton.edu*

During her job talk nine years ago, KMG presented her work on individual differences in (categorical) syllogistic reasoning (Galotti et al. 1986). Faculty present nodded approvingly at her introductory remarks, in which she gave the standard reasons about syllogistic reasoning being important to study because it arguably underlies most or all of the reasoning that goes on in everyday life (Johnson-Laird & Byrne [J-L & B] make similar arguments in their introductory remarks). Undergraduates in the audience were often less sanguine, and at least one would always ask *why* psychologists were so fascinated with syllogisms, as they did not seem to have much to do with "real-world" thinking.

The smug answer to such questions, of course, is to recast the question into a syllogism, as in: "You are claiming that (1) All syllogisms are things that are irrelevant to everyday life. (2) All things that are irrelevant to everyday life should not be studied by psychologists. From these premises, you are concluding that (3) All syllogisms should not be studied by psychologists. But (ha, ha), don't you see that your argument is, in fact, a syllogism?" Such an answer typically gets the speaker off the hook and makes the questioner feel sheepish. Nonetheless, KMG never had the feeling that she had convinced any of the students, merely that she had outfoxed them. We argue now that showing that an argument can be *recast* into a syllogism or other similar form does not at all warrant the conclusion that the reasoner originally reasoned that way.

The undergraduate's question can be framed more broadly: What relation do formal reasoning tasks (such as propositional, syllogistic reasoning, or analogical reasoning tasks) have to everyday reasoning, such as diagnosing a car's failure to start or figuring out how to solve a family budget crisis?

Much of the existing literature centers on *formal* reasoning tasks – those that present all the premises, are self-contained, typically have one correct answer, and typically have established methods of inference that apply to the problem so that it is clear when the problem has been solved (Galotti 1989). The deduction tasks discussed by J-L & B clearly fall into this category. In contrast, everyday reasoning tasks are not self-contained, typically do not present all of the premises (in fact, much of the problem is often in figuring out what the relevant premises are); typically they have no established methods of inference to apply, and may have no solution, one, or many possible solutions. The literature on everyday reasoning is scant (e.g., Rogoff & Lave 1984; Scribner 1986) and appears to be hard to incorporate into existing models of reasoning, at least without considerable handwaving.

J-L & B update and "flesh out" the mental-model approach to deduction. Admirably, they demonstrate how this approach accounts for performance on a variety of deduction tasks. That the model applies to more than one deductive reasoning task is a substantial victory over previous theories of individual reasoning tasks. J-L & B contrast their "models" approach with a "rules" approach that describes deduction in terms of mentally represented inference rules (e.g., Braine et al. 1984; Cheng et al. 1986; Rips 1983); they also set up critical tests of the two approaches. The data they present consistently support the models approach. Thus, their arguments for the greater explanatory adequacy of the models approach over rules approaches in explaining deduction is compelling.

The models approach has another advantage over the inference rules approach: the apparent ease with which it might be extended to everyday reasoning tasks. Although J-L & B do not explicitly address these extensions, it seems easy enough in principle to make them, especially once we have an account of what everyday reasoning is. However, just because the models approach *might* be extended to everyday reasoning *in principle*, it does not follow that it should be or can be.

If the processes responsible for deduction underlie much of everyday reasoning, then the models approach strikes us as clearly "one-up" on the various rules approaches, even if its promise is at this point an uncashed check. If, on the other hand, deduction involves mental processes very distinct from other forms of thought, then narrowly applicable inference rules may be appropriate.

One could argue that deductive inference rules function in reasoning much as syntactic rules function in sentence comprehension: as a description of the operation of an informationally encapsulated module (Clifton & Ferreira 1987; Fodor 1983), the output of which is then combined with other information to yield the behaviors we see. [See also multiple book review of Fodor's *Modularity of Mind*, *BBS* 8(1) 1985.] The analogy may be fine, as far as it goes. However, by this argument, syntactic rules explain a great deal of naturally occurring behavior. We see examples of syntax every time a speaker speaks. How often are propositional, syllogistic, relational, or other kinds of deductive inference drawn in day-to-day living? (And no fair counting the day-to-day inferences drawn by philosophers of science, who are argued by some to "use modus tollens to reason their way to the conclusion that it's time for a cup of coffee" [Pullman 1991, p. 11]). The proposal for inference rules is problematic: If they are only meant to describe certain kinds of thought, they ought to be describing thought that occurs with a moderate degree of frequency.

Furthermore, in formulating and determining the output of syntactic rules, linguists are trying to match real human behaviors (i.e., intuitions). In contrast, the output of deductive inference rules is fixed: They must yield deductively valid inferences. Thus, inference rules do not account for everyday behavior, such as people's deviations from valid inferences, or reasoning in situations that cannot be recast as deductions.

We need to wonder why we are studying deduction so thoroughly. To justify the attention deduction receives, we need to follow one of two paths: Either we apply postulated deductive processes to other forms of reasoning, or we develop some detailed account of the role deduction plays in other forms of reasoning.

## A number of questions about a question of number

Alan Garnham
*Laboratory of Experimental Psychology, University of Sussex, Brighton BN1 9QG, England*
**Electronic mail:** *alang@epvax.sussex.ac.uk*

One of the most frequently asked questions about mental models is: How many (significantly) different models of a given situation are there? This question is important because the theory claims that when a problem requires the consideration of more than one model, it will be difficult. One domain in which strong claims are made about numbers of models is syllogistic reasoning. However, people frequently complain that they find the basis for such claims, as set out, for example, in Chapter 6 of *Deduction*, unsatisfactory. In what follows, I provide a rational reconstruction of the method for deciding how many models a syllogism has, and I show that it has some unforeseen consequences.

The number of models for a syllogism depends crucially on the models of the premises, the initial way those models are combined, the way the integrated model is modified, and how conclusions are read off integrated models. The initial models of the four types of premise (A, I, E, O) in the version of the theory presented in *Deduction* (in minimal form) are:

| A. | | I. | | E. | | O. | |
|---|---|---|---|---|---|---|---|
| [*a*] | [*b*] | *a* | *b* | [*a*] | | *a* | |
| . . . | | . . . | | | [*b*] | *a* | [*b*] |
| | | | | . . . | | | [*b*] |
| | | | | | | | . . . |

Each line in a model represents a type of individual. The method for combining models of the two premises of a syllogism is not stated explicitly, but it appears to be as follows. First, the number of *b*'s in the two models is made the same, by duplicating lines where necessary. Since the models for A, I, and E have only one type of *b*, and the model for O has two, there is never any ambiguity about how to duplicate. The *b*'s in the two models are identified, and the implications of the square bracket notation (exhaustive representation) are worked out for the integrated model.

A conclusion is read off from the combined model according to the three constraints set out in Chapter 2 of *Deduction*. For syllogistic conclusions the only important point is that universal conclusions are stronger than, and therefore preferred to, particular ones. A conclusion is valid if it is not refuted by any other model that is consistent with the premises. For a one-model problem it might be assumed that there is no other model (by definition). However, this assumption is incorrect. If models are differentiated by the type of individual they contain, there are, for example, four types of individual ($+a+b+c$, $-a+b+c$, $-a-b+c$, $-a-b-c$) and hence eight $\left(\sum_{i=0}^{3} {}^3C_i\right)$ explicit models compatible with "all A are B," "all B are C" (assuming existential import). However, not only are all of these models compatible with the conclusion "all A are C," that conclusion is favored by the pragmatic principles in all of the models. A "one-model" problem is, in reality, one in which all models favor the same conclusion. The idea that "one model" problems have more than one model is a welcome one, because it suggests that belief bias effects on one-model problems might be explained in the same way as for multiple-model problems.

Turning to these problems, "All B are A," "No B are C" (*Deduction*, pp. 121ff) generates an initial model of the form:

| | | |
|---|---|---|
| [*a* | [*b*]] | |
| | | [*c*] |
| . . . | | |

which is compatible with several conclusions, but which favors "no A are C." However, revision produces a model that is incompatible with this conclusion. For this reason the revised model is held to be a different one. Thus the classification of this problem as a multiple-model problem depends on which model is produced first, and on the fact that "some A are not C," which is compatible with this model (obviously, since it is the valid conclusion), is not the favored one. Of the alternative models (see *Deduction*, p. 122), the first is compatible with all of "some A are C," "some C are A," "some A are not C," and "some C are not A." The second favors "all C are A." The pragmatic principles do not distinguish between the four conclusions from the first alternative model. Direction of reading might favor A-C conclusions, but this syllogism is not in a figure that produces a figural effect, so there are no grounds for suggesting a direction of reading effect. If this model were generated first, and the conclusion "some C are not A" were derived, this syllogism might be classified as one model on the grounds that no revision produces a model with a favored conclusion inconsistent with the one derived from the original model. To avoid this conclusion it might be suggested that the three models can be distinguished on the grounds that they favor different conclusions, but that would leave open the question of why the number of models, in this sense, affects processing. One possibility is that, rather than just checking that their initial conclusion is compatible with subsequent models, reasoners derive the favored conclusion from every model they construct.

An anomaly in the system of *Deduction* is the existence of two-model problems with valid conclusions. These four problems, like the three-model problems, all have particular negative conclusions (the first syllogism on the last page of Table 6.1 in *Deduction* is really a one-model problem) but, unlike the three-model problems, they also have one particular negative premise. The models shown in Table 6.1 for these models warrant particular affirmative conclusions and are therefore incorrect. These conclusions are ruled out by the "missing" model (which favors "no A are C," a conclusion that is compatible, on the standard interpretation, with "some A are not C"). This model "corresponds" to the first model constructed for three-model problems. It is missed for the two-model problems because the particular negative *premise* is not represented by a model in which, for example, "no A are B," and because the procedure that revises models only adds things and never subtracts them. The simplest way of ensuring that the right conclusions are generated for these syllogisms would be to revise this procedure and to reclassify these problems as three-model ones. The relative ease with which they are solved would then be unexplained.

Indeed, given that syllogisms with valid conclusions divide, in this view, into one-model and three-model problems, this binary classification cannot explain much of the variance in their difficulty. So, even though the mental-model theory is the best theory of syllogistic reasoning we have, it may need a different account of what makes a syllogism difficult.

## Rule systems are not dead: Existential quantifiers are harder

Richard E. Grandy

*Cognitive Sciences Program, Rice University, Houston, TX 77251*
**Electronic mail:** *grandy-a@ricevm1.rice.edu*

Logic teachers are, of necessity, psychologists, although typically they are amateurs, and I am no exception. The most evident amateur conclusion I draw is that reasoning is difficult and the rules far from natural. This accords with the claims of Johnson-Laird & Byrne (J-L & B). A second, almost as evident conclusion, is that existential quantifiers pose the greatest difficulty. If I could find a way of presenting existential quantification that was as graspable as universal quantification my teaching evaluations would rise significantly.

Thus, I disagreed with J-L & B's claim that there is no difference in difficulty between the universal and existential quantifiers. On logical grounds alone, one would expect a difference. Disjunctions are harder than conjunctions, a fact readily observable in the classroom. But whereas universal quantifiers correspond to conjunctions, existential quantifiers are themselves disjunctions – true if any of their (suitably specified) instances are true.

On the general issue of rules, I am uncertain of the exact depth of my disagreement with J-L & B. I have been able to discern three arguments against formal rule systems as representations of actual deduction. The first rests on the evidence in the classroom and the laboratory that humans do not naturally reason using either axiomatic Hilbert or natural deduction systems. However, these are only two kinds of formal rule systems. In my classes, after teaching a natural deduction system I then prove completeness by what amounts to introducing another formal system. This system of rules involves the systematic search for a model which renders the premises of an argument true and the conclusion false. It either produces a demonstration that this cannot be done and hence the argument is valid, or else it produces a description of a model that shows the invalidity of the argument. (The description may be infinitely long.)

The method of semantic tableaux, another variation on this

approach, is incorporated in a number of texts as the systematic formal search for a counterexample as a system of formal rules. As we would all expect, the difficulty of the search increases when disjunctions of possible models are involved, and as I and most logicians would expect, the difficulty also increases when existential quantifiers are present. These particular sets of rules were chosen as much for formal goals as for psychological ease, so I am not arguing that they are exactly the right representation of ordinary deductions, but I do believe that J-L & B have not presented a conclusive case against formal rule systems generally.

The second argument is the alleged suppressibility of *modus ponens*. Given a premise "If she meets her friend, Mary will go to the play" and "Mary meets her friend" the consequent will be deduced by *modus ponens* by most deducers. However, J-L & B have found that if they also present a second premise, "If Mary has enough money, she will go to the play," that reasoners will not draw the conclusion that Mary goes to the play given that she meets her friend. The authors conclude that *modus ponens* has been "suppressed" and thus is not a mental rule. Perhaps, given the second premise, subjects mentally rewrite the first premise as "If Mary meets her friend and she has enough money, she will go to the play," in which case *modus ponens* is not suppressed but is inapplicable.

The third argument is related to the issue concerning existential quantification. J-L & B's argument that existential quantification is no harder than universal seems to have two bases – one a conceptual analysis and the other an experimental one. On page 136 they give an example of a derivation using universal quantifiers, noting that a comparable problem with existentials "differs only in that the existential quantifier, 'some', in the second premise has to be existentially instantiated, and so the quantifier restored at the end of the derivation is also existential. There is no principled way in which the derivations for the two sorts of problems can be made to differ in length."

There is no recognition that existential instantiation in many systems requires a new subproof, and that in others it requires flagging a variable or in other ways giving special status to the formula in question. (In fact, in the universal derivation there is no mention of the necessary restriction on universal generalization.) This raises doubts in my mind whether J-L & B have a sufficient grasp of what is involved in formal existential inferences.

J-L & B's experimental evidence involves two pairs of sentences. The first sentence of each is "None of the painters is related to any of the musicians," while the second sentences are, respectively:

> Some of the musicians are related to all of the authors.
> All of the musicians are related to some of the authors.

The authors report that subjects drew only 23% correct conclusions from the second pair but 64% from the first pair. They apparently conclude "Hence, there is no intrinsic difference in difficulty between existential and universal quantifiers" (p. 142), but I think that they mean to argue that the difference in difficulty cannot be explained by a difference between universal and existential quantifiers because each problem contains the same number of each quantifier. This is true, but it overlooks the fact that some proofs are much more difficult than others because of the ways in which the quantifier rules interact. In some cases the restrictions can prove a major obstacle to unsophisticated reasoners.

In any event, it is impossible to tell from their description what is transpiring because we are not told in the case where only 23% correct conclusions were reached whether the other subjects mistakenly thought no inference could be drawn or if they drew incorrect inferences. My own bet would be on the latter, since most untrained subjects have no intuitive grasp of the restrictions on quantificational inferences. Indeed, most trained subjects lose their grasp fairly quickly if they do not rehearse, and there have even been logic texts which got the subtleties wrong!

## Mental models: Rationality, representation and process

D. W. Green

*Department of Psychology, Centre for Cognitive Science, University College London, London WC1E 6BT, England*
**Electronic mail:** *d.w.green@ucl.ac.uk*

It is a pleasure to read Johnson-Laird & Byrne's (J-L & B's) *Deduction*. It marshals the arguments and evidence for a mental-model theory of deduction with sustained clarity, force, and wit.

***Hybrid rationality?*** Like theories based on mental rules, the theory of mental models proposes that there is a general competence to be explained. The arguments and experimental evidence favour the mental-model account of this general competence over a rule-based one. But is model construction and manipulation necessary for correct deductive performance? The short answer is "no." Trivially, if the answer to a problem is known, it can be retrieved. More pertinently, as J-L & B acknowledge, some individuals, tutored in logic or argumentation, may use rules or "tricks" for certain tasks. Different forms of reasoning may therefore coexist within the same individual. In addition, individuals may find shortcuts to solve specific kinds of problem. What was derived initially by envisaging a model might, during the course of the experiment, result in procedures which derive answers directly from the linguistic content. Hence, there are a variety of circumstances where model construction need not mediate rational response. If this conclusion is granted, it points to the need to consider individual patterns of performance.

Despite individual differences, I imagine that J-L & B would wish to claim that human rationality is fundamentally based on a unitary underlying competence and is not hybrid in the sense of involving both general procedures (e.g., those proposed in the theory of mental models) and domain-specific procedures, such as pragmatic reasoning schemas (Cheng & Holyoak 1985) or the cheater-detector algorithm in the social contract theory of Cosmides (1989; see also Gigerenzer & Hug 1992). Mental-model theory is, of course, more general than any domain-specific theory and is more parsimonious than any hybrid account; but neither of these properties precludes the psychological possibility that specific procedures are invoked in particular domains. It is not sufficient to show that certain findings claimed as support for domain-specific procedures can be explained *post hoc* by the theory. From an experimental point of view, more refined performance measures are required to contrast the predictions of model theory with those of domain-specific accounts of domain-specific problems, that is, of problems for which the theory of narrower scope is suited. Alternatively, empirical work on mental models could be extended to include neuropsychological data (e.g., studies of individuals with damage to the frontal lobes) that might reveal any functional dissociations (see Shallice, 1988; see also multiple book review, *BBS* 14(3) 1991), and thereby enrich the debate on the nature of the underlying cognitive architecture mediating reasoning performance. The work of Leslie and others on autism (e.g., Leslie & Thaiss 1992) confirms the possibility of dissociations in central processes.

***Representational form.*** The procedures of model theory can be viewed as basic cognitive operations that allow the construction of models in a variety of representational forms (e.g., visuospatial). Although J-L & B rightly focus on the structural characteristics of models, it is natural to wonder about the form

in which models are represented mentally and indeed, such representations need to be specified if complete computational descriptions are to be given of performance on specific tasks. We can gain some clues by looking more closely at the process of model construction. This process treats the propositions expressed as data and constructs a mental world in which these propositions are true. Understanding is tied, temporarily at least, to the acceptance of the truth of a proposition in a way compatible with Spinoza's conjectures (see Gilbert 1991) and consistent with the way that perceptual input guides action. If there is a close relationship between thought and perception, one might expect to find correlations between performance in a perceptual domain and in a reasoning domain. Yet, as far as I know, such correlations are not obtained. Once again, neuropsychological data might prove informative. For example, subjects with deficits in visuospatial processing should perform more poorly on problems involving spatial descriptions but should not necessarily fail on syllogisms that do not reference a spatial dimension.

***Processing the model.*** A robust finding is that performance is worse on problems that require subjects to consider more than one model. By itself, such a finding is open to two interpretations. Subjects may stop reasoning when they reach a conclusion or they may seek to envisage alternative models and fail, perhaps because of working memory constraints. In some studies, the former interpretation seems to be correct (Lee & Oakhill 1984), whereas in others (e.g., Johnson-Laird & Bara 1984), the latter interpretation seems to be correct. A crucial question, as J-L & B recognize, concerns what factors cue subjects to construct alternative models or to flesh out their initial model. They identify a number of cues, namely: The meaning of the premises may permit different initial models; initial conclusions may be considered unbelievable; the tokens depicting particular entities may be represented as not exhausting the set of such individuals. In addition, I imagine that some subjects invoke a heuristic: "Search for counterexamples." Given the variety of possible cues, it seems unlikely that there is a single psychological algorithm for evaluating conclusions. In this view, the proposed algorithm (p. 182), which first negates the conclusion and then sees whether there is an alternative model of the premises consistent with it, is one of a number.

Given the above, it seems desirable to obtain more direct evidence about the process of fleshing out the model in specific tasks so as to develop more complete accounts. In fact, a recent study which required individuals to externalize their thinking under different constraints (Green 1992) confirms the core of the mental-model account of performance of the selection task. It has also revealed an apparent paradox. Some individuals envisaged the critical counterexample but failed to select it. Such a finding implicates a postdeductive process which evaluates possible selections.

## The logical content of theories of deduction

Wilfrid Hodges

*School of Mathematical Sciences, Queen Mary and Westfield College, University of London, London E1 4NS, England*
**Electronic mail:** *w.hodges@qmw.ac.uk*

Johnson-Laird & Byrne's (J-L & B's) book argues that we make deductions not by applying rules of inference to representations of the logical forms of our premises but by a process which involves building mental models of the premises and searching among them for counterexamples to the conclusion. Experiments are reported which (it is claimed) support this theory.

Let it be said at once that the mental-model theory of deduction has a pictorial quality which many people have found appealing and inspiring. Nevertheless, J-L & B's book falls short of the standards one would expect on logical writing today. There is a fair amount of symbolism, suggesting precision, but most of it is so poorly explained, or so loosely attached to the matter in hand, that the reader can only guess what is meant; time after time it happens that an interpretation which works on page *X* won't work on page *Y*.

From dozens of examples I choose two which are central. The first is an explanation of how we carry out *modus ponens;* that is, given "If $p$ then $q$" and "$p$," how we deduce "$q$" (p. 47, repeated on p. 196). It is claimed that we start with the first premise, forming two mental models; the first model represents the case that $p$ and $q$ hold, and the second "has no explicit content." The second premise then eliminates the second model, since it is true already in the first model. Finally, from the first model we read off $q$. It is hard to believe that this protocol has any logical connection with the deduction that it is supposed to perform.

The second example is the notation "$[[a]b]c$" which appears on page 121 in the treatment of syllogisms. It is said to signify "that $a$ is exhausted with respect to $b$, and $b$ is exhausted with respect to $c$." The notion of being "exhausted with respect to something" is not explained in the text and it means nothing in logic; I dare wager it means nothing in psychology either. The interpretation which comes first to mind is that the notation means "All $a$'s are $b$'s and all $b$'s are $c$'s"; but unfortunately this reading implies that in order to use the model, we already have to be able to carry out exactly the deduction which the model was intended to explain.

This makes it impossible to comment in detail on the theory proposed in the book; I simply do not know what that theory is. Two points of methodology call for some remarks, however.

The first is the way in which J-L & B pose the basic contrast between the formal rules theory and their own mental-model theory. Supposedly these are two theories about how our minds work. But the authors tend to explain the difference by using notions from the mathematical theory of formal systems. A typical example is on page 212, where they explain that mental models "do not contain variables." Without some explanation of what it is for a mental representation to "contain a variable," this is meaningless. (My own impression is that many of the mental models described in this book do in fact contain components which behave pretty much like variables, if one looks at variables in the appropriate way.) Because of this mismatch between the phenomena to be explained and the concepts used to explain them, the book fails to establish a genuine difference between formal rules and mental models.

The second point of methodology concerns the claim that a theory of deduction based on mental models "predicts which problems will be difficult and it predicts which errors ordinary individuals will make with them" (p. 131). This claim will not survive a closer look at what is meant by "a theory based on mental models." Take, for example, the case of syllogisms, as in Chapter 6. If the theory in question is either (1) the general theory that we make deductions by forming models of the premises and looking for counterexamples to the conclusion, and so on, or (2) the theory of models of syllogisms as presented in the chapter, then it is too imprecise to have the consequences claimed, for example about the numbers of models needed for each syllogism.

One suspects that the authors may have in mind (3), the detailed theory propounded in Johnson-Laird & Bara (1984). This theory is different from the one outlined in the chapter, but it seems to underlie some of the discussion, and it is precise enough to be written as a computer program. The problem with this third theory is that it involves, among other things, fourteen "principles" for carrying out operations, some of them more *ad hoc* than others. Since the theory has almost as many degrees of freedom as the data to be explained, the reasonable fit is hardly impressive. To justify their claim, the authors need to produce a theory which is precise enough so that the reader can verify what predictions it makes, and one that is also derivable from

the general mental-model theory in a way which is not arbitrary. They have not done this.

Some of us – one might single out Jon Barwise at Indiana and Johan van Benthem in Amsterdam – have been urging for some time that people with an interest in logic should talk to one another and build up a common expertise. We owe it to the next generation to see that this happens across the boundaries of psychology. The work in this book is worth doing properly.

## Architecture and algorithms: Power sharing for mental models

Robert Inder

*Human Communication Research Centre, University of Edinburgh, Edinburgh EH8 9LW, Scotland*
**Electronic mail:** *r.inder@ed.ac.uk*

Johnson-Laird & Byrne (J-L & B) claim that people make deductions by combining premises into composite representations, or models, generating candidate conclusions from these models, and then (unsuccessfully) searching for counterexamples. They suggest that this procedure, which is directly derivable from a (model-theoretic) logician's definition of a valid inference, is the psychological basis of human inference, at least under some circumstances. They contrast it to other approaches based on deriving conclusions by a series of steps, each governed by some kind of inference rule, suggesting that people make inferences without applying inference rules by manipulating mental models which have "a structure that is remote from verbal assertions, but close to the structure of the world as humans conceive it" (p. 207) and which are, in some sense "semantic."

This suggestion is not straightforward, because one can clearly translate any given problem into a suitable logic and devise a (rule-driven) inference system to manipulate logical forms to produce the predicted model construction and transformation. Indeed, J-L & B know this, since Chapter 9 discusses computer programs that do just this. Given that the inferences being made will inevitably be describable by inference rules, what sense can we make of the suggestion that inferences were made without recourse to those rules?

For some kinds of inferences, one can choose representation schemes that ensure that the representations built will embody the results of certain kinds of inferences. We see this when the spatial arrangement of model components for physical layout problems is chosen to reflect the situation being modelled. But this only works for certain limited kinds of inference. A more general possibility becomes apparent when we recognise that psychological theories must postulate both algorithms – what is going on – and the computing engine on which they are executed – what Pylyshyn (1980) has called the "cognitive architecture." A claim that a piece of knowledge becomes available "without inference" amounts to a claim that it is not generated by the algorithm being described but by some part of the cognitive architecture. Moreover, if mental models are represented for manipulation by the cognitive architecture and this manipulation takes place in line with long-term memory and general knowledge of the world, there is a sense in which the models can reasonably be claimed to be semantic.

How relevant information is retrieved from long-term memory is one of the most awesome feats of the human mind. It happens automatically – we seldom forget the need to walk round furniture. And it happens quickly – whatever mechanism is responsible for searching and retrieving from a lifetime's long-term memory is clearly very different from the processes which grapple for seconds with decisions about entities in syllogisms. These are reasonable properties to be exhibited by a process deemed to be supported by the cognitive architecture.

Thus we can make sense of the "semantic" and "without inference" claims with respect to the reasoning process by pushing functionality into the cognitive architecture. Provided it is clear what functionality is being assumed, a theory that offloads work onto the cognitive architecture is as plausible as the assumptions it makes about the functionality available. That the mind is supported by mechanisms for representing a situation and bringing background knowledge to bear on it has immense plausibility; it is, in effect, a suggestion that brains come with "mental-model accelerators" fitted.

However, as Pylyshyn (1980) argued, we must minimise the power of the services that are assumed to be provided by the cognitive architecture. If too much is assumed of the architecture, our task theories become vacuous – they are simple, but we are left with disconcerting lumps under the carpet of the cognitive architecture.

The simplest form of model would represent a single state of affairs, and indeed Johnson-Laird has always suggested that mental models share the structure of the situation being modelled. However, J-L & B extend models to include negation and indications of "set exhaustion." This appears straightforward if, as in their examples, the property being either denied or "exhausted" is a primitive – that is, a "word," in either English or Mentalese. Restricting properties which involve combinations of others requires handling scoped logical operators. Both run contrary to J-L & B's notion that "A model . . . has a structure that is remote from verbal assertions." Moreover, these mechanisms allow one model to represent many states, and if multiple negations within a single state are not prohibited, can give almost unlimited representation of uncertainty. Not only does this greatly increase the power of the model-manipulation facilities being postulated within the cognitive architecture, it also undermines the whole basis for identifying what a model represents. We lose our intuitions about what a model is – which matters immensely, since J-L & B never actually define it. Thus, counting the models required to represent a set of cases – "a single model for a single situation" (p. 196) – becomes merely a consequence of the precise details of the logic that is chosen.

In seeking to establish mental models as a useful theorising tool, J-L & B have created a complex representation with many task-oriented features and have thus tacitly called for highly sophisticated processing from the cognitive architecture they assume. The result is a seductively elegant account of reasoning, but one that needs powerful reasoning machinery to provide the complex task-oriented operations it uses. J-L & B must reduce the demands they make on the cognitive architecture. This may well make their task theories more complicated, but this would be a good thing: It is the effect of pulling the complexity of the task into the theory of the task, where it can be seen. Inder (1987) shows that it can be done for syllogisms. The aim for all who would base a cognitive theory on mental models must be to show how they can support task-oriented reasoning without subsuming it.

ACKNOWLEDGMENTS

The support of the Joint Councils' Initiative in Cognitive Science and HCI through project number G9018050 ("Signal") is gratefully acknowledged. The Human Commuuication Research Centre is an interdisciplinary research centre funded by the Economic and Social Research Council of the United Kingdom.

## The content of mental models

Paolo Legrenzi[a] and Maria Sonino[b]

[a]*Department of Psychology, University of Trieste, 34123 Trieste, Italy and* [b]*Department of Psychology, University of Padua, 35100 Padua, Italy*
**Electronic mail:** *girotto@univ.trieste.it*

Johnson-Laird's theory of mental models predicts that erroneous inferences correspond to descriptions of a subset of the

models of the premises of the problems to be solved. It is intuitive that inferences calling for fewer mental models are easier than inferences calling for more models, as working memory has a limited space. By contrast, it is not clear why some subjects stop building the necessary mental models, being satisfied with an incomplete representation, while other subjects are able to generate alternative models and, consequently, avoid errors. One possibility is to attribute these differences to different capacities (or to differences in mechanisms of engagement) of working memory of different subjects.

If we take this position, relating the numbers of models to the load on working memory, we can try to extend this explanation to the results obtained in reasoning experiments in which the same logical structure is filled with different contents. In these cases, we have to explain difficulties and errors not only on the basis of the quantity of models – more models, more load – but also on the basis of their quality – more experience, less load.

We know that realistic conditional rules, such as "If a person is drinking beer, then the person must be over 18" make Wason's (1966) selection task easier to solve. According to the model theory, experience with the rule about beer drinking helps to flesh out the models with more explicit information:

| | |
|---|---|
| [drinking beer] | over 18 |
| not drinking beer | [not over 18] |

. . .

and so subjects will tend to select the card corresponding to the negated consequent. This result is consistent with the general hypothesis that reasoners focus too exclusively on what is explicitly represented in their initial models and so they overlook alternative possibilities. In fact, the initial models of the conditional rule make explicit those items mentioned in the rule; this explains why, in the standard selection task, it is difficult to take into consideration the cases corresponding to the negation of the consequent. This explanation implies that, when the subjects are able to take into consideration the negation of the consequent, we have to explain this facilitation in terms of contents or contexts that make it easier to avoid overlooking alternatives. In particular, mental-model theory explains why the right selections become easier as a result of a variety of experimental manipulations, such as the use of a deontic content, a linguistically simpler rule, and so on. In conclusion, the model theory can explain better than a traditional hypothesis, such as Evans's (1989) matching bias, the results obtained through the various experimental manipulations affecting contents.

Nevertheless the theory seems incomplete in relation to the explanation of the interaction between content and ease of building mental models. In fact, only a posteriori does it seem possible to say that certain types of content (for example, deontic rules) help subjects in fleshing out models in specific domains, such as the search for transgressions of rules. The role of certain linguistic transformations, (such as the "only if" transformation) is evident, as they help to elicit negative information. Less evident is the role played, *ceteris paribus*, by the different contents of mental models. The theory is silent about the specificities of these facilitations unless they are put in relation, as usual, with the load on working memory. In this last case, however, it does not seem so clear to us why, for example, subjects' experience with deontic rules together with the "only if" form of conditional is of relevant help in fleshing out models. Extending and specifying the theory in relation to the content mechanism will be very useful to avoid circularity (e.g., facilitation, explained as ease in fleshing out the models, only when facilitation is obtained).

## Gestalt theory, formal models and mathematical modeling

Abraham S. Luchins[a] and Edith H. Luchins[b]

[a]*Department of Psychology and* [b]*Department of Mathematical Sciences, Rensselaer Polytechnic Institute, Troy, NY 12180-3590*

We believe that the thesis presented in Johnson-Laird & Byrne's (J-L & B's) important book is partly – but not completely – in keeping with the views of Gestalt theorists and, in particular, with those of one of its founders, Max Wertheimer (1880–1943).

In his 1925 paper on the syllogism and productive thinking, Wertheimer criticized the traditional syllogistic process as often "empty, inadequate, and sterile" (translation in Ellis 1938, p. 274), rarely yielding what was not already explicit in the premises, seldom resulting in genuine, productive thinking (see also Duncker 1926; 1945; Luchins & Luchins 1965b; 1970; 1991; Wertheimer 1945). Wertheimer, who taught psychology and logic, maintained – as J-L & B do – that the formal laws of logic are not the laws of thinking. In his seminars at the New School for Social Research (Luchins & Luchins 1970; 1987; 1991), Wertheimer criticized Piaget's views on children's thinking. He would be opposed to the notion that "reasoning is nothing more than the propositional calculus itself" (Inhelder & Piaget 1958, p. 305; quoted on p. 23 by J-L & B). J-L & B reject relativistic views of logic (p. 209), as did Wertheimer (1912), who eighty years ago wrote about primitive peoples' logical and numerical concepts and thought.

Wertheimer did not claim that there is no faulty reasoning or that an apparent invalid inference is necessarily a valid inference from other premises, although in recent years Henle (1962) has defended such views (J-L & B, p. 18). With his conception of people as *Homo sapiens*, rational beings who can be blinded or deceived by internal and external factors, Wertheimer would probably agree with J-L & B's belief that "people are rational in principle, but fallible in practice" (p. 19).

Like J-L & B, Wertheimer recognized that thinking was affected by the context, content, and meaning of the premises and not just by the logical form. He wrote, "On truth" (1934) and criticized traditional logic for not being concerned with truth; he would welcome J-L & B's references to truth and truth-preserving properties (p. 214). Their contention that the formal model fits the structure of the situation is in keeping with the Gestalt (Wertheimer's) notion of isomorphism (Köhler 1938). Wertheimer sought to develop a Gestalt logic that would take into account context, meaning, truth, and fit, and would not be piecemeal and summative in nature (Luchins & Luchins 1965b; 1970; 1991). Would he therefore have embraced J-L & B's formal model theory? We think not, although he would undoubtedly have rejected the rule theory.

There is an underlying "and-summative" quality in *Deduction*. Reasoning is divided into calculation, deduction, and induction (p. 2), with creation and association added (p. 193). But ordinary people do not divide thinking in these ways. Moreover, an individual's testing of models and search for alternative models are described as involving alterations of different features in a piecemeal manner. Furthermore, the approach is bottom-up with the ultimate aim of cognitive science being to combine "the underlying components out of which deduction and other cognitive abilities are assembled" (p. 214).

We are disappointed at J-L & B's failure to refer to the Gestalt psychological literature on thinking and perception. We applaud the ingenious experimental approach but note that it is now largely limited to artificial situations, although there is a promise to deal with real-life situations. We think that the importance of deduction is overestimated. Even in mathematics, the most deductive of the sciences, deduction tends to play a lesser role in discovery than in the final presentation which may be (but need not be) in hypothetical-deductive form. We do not agree that "human beings draw parsimonious conclusions for

themselves" (p. 183). They can at times draw and maintain redundant and contradictory conclusions (Luchins & Luchins 1964; 1965a). J-L & B have overestimated an untrained individual's ability to formulate and manipulate a model, to draw conclusions from it, to validate it, to look for counterexamples, and to search for alternative models. This is virtually a description of mathematical modeling (Giordano & Weir 1985), which requires training, even on the part of professional mathematicians.

An NSF-supported, ten-day workshop on teaching mathematical modeling was held at West Point this summer for college professors of mathematics. Starting with a real-life problem or a set of data as observations (call it the target), one looks for a mathematical model that seems to hold for the target, or for a simplified version of it, implements it, perhaps by model-simulation, draws conclusions from the model, compares them with the target, studies how good the fit is, refines the model, or looks for another more suitable model. In some cases the target points to the mathematical model. For example, it may suggest exponential growth. But in other cases even the experienced mathematical modeler may not have a clue as to an appropriate model and may have to decide on trying a linear model or a nonlinear model (there may be infinitely many nonlinear models), a continuous or discrete model, a deterministic or stochastic model, and so on. It may be impossible to exhaust all the feasible models. There are even different models to use in determining how good is the fit of the model; mathematical modeling is not generally a simple process. However, when it comes to deduction, the book depicts the unsophisticated reasoner as capable of having a formal model that more or less fits the situation, of validating conclusions from it, even of looking for counterexamples, and of shifting to other models. To test a conclusion may require the reasoner to exhaust all models that fit the premises. It seems that Johnson-Laird & Byrne see everyone as a natural mathematical (or perhaps nonmathematical) modeler. Perhaps it is so in the best of all possible worlds.

ACKNOWLEDGMENT

This commentary was written while both authors were visiting the Department of Mathematical Sciences, United States Military Academy, West Point, NY 10996-1786.

## Visualizing the possibilities

Bruce J. MacLennan

*Computer Science Department, University of Tennessee, Knoxville TN 37996*

**Electronic mail:** *maclennan@cs.utk.edu*

I am in general agreement with Johnson-Laird & Byrne's (J-L & B's) approach and find their experiments convincing; therefore my commentary will be limited to several suggestions for extending and refining their theory.

***Images and models.*** The distinction between *models* and *images* is treated briefly in *Deduction* (pp. 38–39, 93, 99–100, 140), but four differences are described in Johnson-Laird (1983, especially Ch. 8). I will argue that the distinction is better treated as a matter of degree rather than of kind.

First, Johnson-Laird (1983, pp. 157, 165) defines images as models from a particular viewpoint, but this is inessential to the idea of an image. For example, the transformation of an image from an oculocentric frame to an object-centered frame is just one of many transformations it may undergo in being put into more abstract form. It seems arbitrary to treat differently images in different reference frames, because many of the same processes (e.g., rotation and translation of components) will be applicable to both.

Second, it is claimed that models differ from images in that models can represent negation and disjunction whereas images cannot (*Deduction*, pp. 38–39, 196; Johnson-Laird 1983, pp. 423–24), but it is better to consider these "propositional tags" to be related to *intentions* toward perceptual and motor images. Here, by "intention" I mean a functional relation to a component of a mental representation (including both images and models). Thus it has both *form* ("anticipation that," "denial that," "surprise that," etc.) and *content* (indicating its object); essentially a predicate plus a vector. The point is that intentions toward sensory images are closely related to intentions toward mental models.

For example, orientation toward the absence of an expected object is an intention, the content of which is the absent object. Thus, for perceptual images that are sufficiently abstract, there is a mechanism for representing the negation of a token within an image. Similarly, the presence of an unexpected object can produce an orienting reaction and generate an intention of the form "this shouldn't be here." Intentions toward absent and unexpected objects are closely related to negations of components of mental models, which are intentions of the form "this can't be here." Other "tags" proposed by J-L & B, such as "exhaustive representation" (p. 45), are intentions corresponding to perceptual intentions, such as those of the form "this is typical" or "this must be here."

Furthermore, just as we may judge an entire scene beautiful, threatening, or absurd, so an entire mental model may be the content of an intention to the effect that the entire model is impossible, incoherent, or unacceptable; this is J-L & B's negation of an entire model, but it corresponds to intentions referring to an entire image. Disjunction is not a relation that has to be represented within images, since a disjunction of models is represented by multiple models in working memory (e.g., p. 52), and this works as well for images.

The third distinction between models and images is that the tokens of a mental model may not be accessible to consciousness (*Deduction*, p. 39), whereas, presumably, the tokens of an image are. These "invisible tokens" may simply correspond to unattended elements in a perceptual image; that is, they are represented in the background, but not the object of an intention. For components of both images and models, presence in conscious awareness is a matter of degree, with some elements being more salient because they are the objects of intentions. Although JL & B (p. 39) say, "What matters is, not the phenomenal experience, but the structure of the models," a consideration of the phenomenal experience may benefit a more general understanding of mental representation.

Finally, J-L & B cite as evidence in favor of models over images that there was no significant difference in the performance of subjects on relational reasoning problems that differed in imageability (p. 140); but this is not supported by the experiments described, because all the relations they cite are conducive to visual reasoning. The relations "in the same place as" and "equal in height to" have obvious visual representations, and "related to in the simple consanguineal sense" is simply visualized as "in the same place as." Experiments to refute imageability are in fact hard to design, because 3D space is so powerful a medium for relational reasoning. On the other hand, positive evidence for imageability comes from the results presented on p. 97: There was no significant difference in the performance on two-dimensional and one-dimensional problems. This suggests that we use our two-dimensional visual reasoning ability for both one- and two-dimensional situations, further evidence that models are abstract images.

In conclusion, the difference between images and models is not one of kind, but a matter of degree of abstractness: Models correspond to images at very abstract stages in the sensorimotor circuit, where we find abstract reference frames, intentions of various kinds and a continuum of degrees of presence to con-

sciousness. Treating models and images as two species of the same kind may illuminate both and, in addition, expose the nature and role of intentions in cognition.

***Cultural universals.*** J-L & B (pp. 207–9) claim that the culturally universal aspect of rationality is "the search for counterexamples." I suggest that this should be generalized as follows: (1) The function of comprehension is construction of an acceptable model of the stimuli, which is more than a consistent model in that it must be acceptable within a cultural context, and less than a consistent model in that it may contain culturally permissible inconsistencies. (2) The function of validation is to search for unacceptable models that would cause a hypothesized conclusion to be rejected. Consistency is not a universal logic, even within Western culture (Prier 1976).

***Comprehension and connectionism.*** J-L & B describe deduction as a three-stage process, comprising comprehension, description, and validation (pp. 35–36). Comprehension is a kind of constraint satisfaction: finding the best (most acceptable) representation of the input. Connectionism suggests a mechanism for comprehension, because background knowledge and the stimuli define an "energy surface" with multiple local minima corresponding to acceptable interpretations of the stimuli. The interpretation chosen is the global minimum, but if it later becomes unacceptable, the state (interpretation) can rapidly move to the next best minimum. The possible interpretations are, in effect, constructed in parallel, therefore any necessary reinterpretation is more efficient (cf. multistability in perception). Multistability may also play a role in validation, because it provides a mechanism for generating alternate interpretations of the stimuli against which a hypothesized conclusion may be tested.

## Models for deontic deduction

K. I. Manktelow

*School of Health Sciences, University of Wolverhampton, Wolverhampton WV1 1DJ, England*

The theory of mental models has revolutionised research into human thinking, but even its converts can find ways in which it can be challenged. My object here, then, will not be to attempt to sink the theory, but to repair a leak. The leak involves deontic thinking. This field has been invigorated recently by a series of studies of deontic forms of Wason's selection task, so this commentary will focus on the treatment of this paradigm, especially in its deontic form, in Chapter 4 of *Deduction*.

The model theory of the selection task derives from Johnson-Laird & Wason's (1970) "Insight" theory. Its reincarnation in *Deduction* (p. 79) brings with it some problems that remain unresolved. One of the problems in the old theory was to account for the difference between the common selections of $p$ alone or $p$ and $q$ in the state of "no insight." In the former case, subjects were said to be adopting an implication interpretation of the target conditional, whereas in the latter they are adopting an equivalence interpretation. "Complete insight" was held always to consist of the $p$ and $\neg q$ selection, but this is only insightful with an implication interpretation: Equivalence interpreters should select all four cards or alternatively change to an implication interpretation somewhere along the way. *Deduction* (p. 80) repeats this elision.

Deontic conditional reasoning has one clear characteristic: People are very good at it. Cheng and Holyoak (1985) were the first to show that a deontic context will facilitate conditional reasoning almost irrespective of the content of problems. Both Johnson-Laird & Byrne (J-L & B) and I (see Manktelow & Over 1991) have tried to invoke the model theory to explain the data, but my argument is that J-L & B's account is incomplete.

Look at what is meant by "good" reasoning. For J-L & B, as for most investigators, "good" selection task performance means the traditional facilitation effect, that is, selection of just the $p$ and $\neg q$ cards (see above). Recent research, however, shows that this is too narrow a definition. Cosmides (1989), for instance, reported that subjects would select the $\neg p$ and $q$ pair when presented with the conditional "If a man has a tattoo on his face then he eats cassava root" embedded in a context portraying getting tattooed as a cost and eating cassava as a benefit. Plainly, something other than the traditional "correct" response is being facilitated here. To explain this, J-L & B (*Deduction*, p. 78) propose that the rule is converted by subjects to "$p$ only if $q$" form. What were $\neg p$ and $q$ now become $p$ and $\neg q$, and all is well with the facilitation effect.

However, since *Deduction* we have demonstrated that it is possible to elicit the $\neg p$ and $q$ selection routinely (Manktelow & Over 1991; 1992a). We don't think there is anything irrational in this behavior: People are still being good at the task. Deontic discourse is concerned fundamentally with expressing subjective utility, and deontic thinking is the act of assessing utilities so as to achieve a goal. Uttering a deontic conditional involves a further element: There are at least two parties. One (the agent) utters the rule, the other (the actor) is its target. For such an utterance to be acceptable, each party must make assumptions about the preferences of the other. Thus when a mother says to her son, "If you tidy your room then you may go out to play," she must assume he doesn't want to tidy his room but wants to go out, that is, he prefers going out to staying in; if he would rather tidy his room and stay in, the rule is vacuous.

Using this kind of rule in a selection task produces *four* potentially correct responses, not one, depending on which party's perspective and action one is considering. Two consist of the $p$ and $\neg q$ selection, two of the $\neg p$ and $q$ selection; all are equally rational in the prescribed contexts.

To explain why these responses are facilitated, and why they are rational, J-L & B have to invoke rule conversion, as they do in explaining Cosmides' results (see Johnson-Laird & Byrne 1992). This is unnecessary, however, if one rejects the narrow reading of correct conditional reasoning in *Deduction* and allows models to represent utilities. There are two further reasons for doing this.

The first arises from the second component of the general model theory of the selection task (*Deduction*, p. 79): Subjects "select those cards for which the hidden value could have a bearing on the truth or falsity of the rule." This won't work for deontic tasks, because there one is charged not with assessing the truth status of the rule, but with detecting possible violations of it. Violations are actions, and there are several ways of doing this for any one rule, as we have seen.

The second concerns the "engine" of deontic thought. Why utter a rule in the first place, and why search through models? Utility is the motivation in both cases: An agent prefers something to be done or not done and manipulates the actor's preferences to achieve it. Both parties are sensitive to deviations by the other and will search efficiently for likely instances, because doing so is in their material interest. This basic semantic point is missing from most psychological accounts of deontic thought and must form part of any complete theory. It is rational to behave in this way, if rationality can concern achieving goals as well as adherence to logic (cf. Evans, in press).

Such is the richness of deontic thought that I agree with Johnson-Laird & Byrne that a model theory is far more likely to capture it than are either formal or content-dependant rule theories. A complete model theory of deontic thought will need to represent not just content tokens, however, but also preferences among them; search procedures will have to acknowledge their motivational and social basis and, I suspect, other factors such as subjective probability. It should be an interesting voyage.

## Situation theory and mental models

Alice G. B. ter Meulen
*Department of Philosophy, Indiana University, Bloomington, IN 47405*
**Electronic mail:** *atm@ucs.indiana.edu*

The inference rules of predicate logic were designed to disregard subject matter, content, context, and the order of the information given in the premises of an argument. That is why predicate logic constitutes such a universally applicable, idealized core of any inferential system, human or artificial. But in reasoning we ordinarily do pay attention to all these features that bear on the information expressed and we process the information in tbe premises in the given order. We also make mistakes, draw invalid inferences, and find some inferences easier than other, equally valid ones. Johnson-Laird & Byrne (J-L & B) assess the discrepancies between familiar logical inference rules and patterns of performance in reasoning experiments with human subjects. As a natural language semanticist, I share a good deal of their concern about these common idealizations in most logical theories and regard inference too as a core cognitive process. Cognitive science should characterize inference in a way that does equal justice to our intuitions on (in)validity of inferences expressed in natural language and to the fallibility of actual human reasoning or artificial inference engines. We need a theory of inferential competence that helps us understand why it may fall short in performance.

On some fundamental points I agree with J-L & B; for example, the three constraints governing inference (p. 21–22) correctly require that a conclusion should not add new information or weaken the given information but must express part of the given information in a new "parsimonious" form. In their complaints about the rules of disjunction introduction and conjunction introduction or elimination one can recognize some of what motivated the current developments in situation theory (Barwise 1989; Barwise & Perry 1983; Barwise et al. 1991; Cooper et al. 1990), discourse representation theory (Kamp & Reyle 1992), and other innovative inferential systems (e.g., linear logic and dynamic modal logic) that focus on information, context, and content. Conclusions should not only preserve the assumed truth of the premises, they should be about the same subject matter and should not affect the context. J-L & B rightly emphasize that the structure of the described situation matters essentially to the inferences people make, though they never say what a situation or its structure is. In describing a situation, the order in which the premises are provided determines the given partial information that is used in reasoning. Most surprisingly, this is just what is ignored in J-L & B's research. Their experimental material avoids pronouns, premises are virtually always stative (e.g., *a is in front of b*) and in present tense. They never use conditionals with the if-clause presented after the then-clause, nor do they systematically vary the order of the premises. Yet their results may well be significantly affected by the order in which the information is given.

I confess I have never conducted any experiments on conditionals, nor do I know of any experimental results on how the order of conditional clauses may affect ease of inference. But I boldly conjecture that the ease of *modus ponens* inferences J-L & B observed, should be significantly decreased when the order of the clauses is inverted. A partial cause of the observed greater difficulty of *modus tollens* arguments may be found in its "backwards" reasoning from the negation of then-clause to the if-clause first presented. With inverted order of the clauses, *modus tollens* inferences may be substantially easier. Experimental results on such issues would really tell us something about how we represent information we are given by changing the context or preserving aspects of it.

Lacking any answer on this order issue, I am not in the least convinced that J-L & B have succeeded in arguing against a rule-based account of inference as a cognitive process. Making *modus tollens* a derived rule increases the length of proofs, but it says nothing about processing complexity. We already know that the rules of predicate logic are merely concerned with truth-preservation and hence are not good candidates for modeling how we reason with partial information about a described situation. Situation theory makes a much better tool-kit for the semantics of natural language. Content, context, situated reasoning about described situations, and dynamic context-change are its core concepts to represent the informative content of linguistic input. But it also makes precise how information describes a situation, what a situation is, and how we reason with partial information about it. In situation theory, conditionals are constraints that make it possible that a situation described by the if-clause can contain information about another situation. For example, *if Joan winked, it is time to leave* expresses that any past situation in which Joan is winking makes this a situation in which we should leave. Barwise's (1986) article on conditionals in situation theory appeared in the same volume as Johnson-Laird's (1986) paper about mental models (Traugott et al. 1986). As one of the editors of that collection, I am distressed to see that six years after that book appeared so little has been accomplished to bridge their views.

There are other points at which J-L & B leave us so much in the dark that I cannot see if any positive gains have been made. Let me review these points succinctly.

***1. What makes these models mental?*** They are claimed to be mental objects, but not images (they work for nonvisual relations too) or products of the imagination. Supposedly subjects construct the same model of the premises, at least individual variation is disregarded in the theory. Models for quantified premises contain "mental tokens" (p. 144), instantiating the universal quantifiers "exhaustively" and not the existential one. What is this instantiation, if not a logical process of substituting a referring expression for a variable or parameter? Do you and I have the same mental model if we make the same inferences? Or does it matter how we arrived at the model, and what other information we use in drawing conclusions? How about me and Putnam's Twin Earthling who believes the same things I do about water, but water is XYZ, not $H_2O$, on Twin Earth. If our models are mental, we must have the same one, but in that case meaning no longer determines reference.

***2. Situations and models.*** J-L & B distinguish implicit and explicit models and "flesh out implicit models" in searching for counterexamples; this is reminiscent of the systematic, alas rule-based but genuinely semantic search for possible falsification in Beth-tableaux (cf. Partee et al. 1990, Ch. 2). But in what sense are their mental models real models of the premises? Logic certainly gave us a clear understanding of what a model is supposed to do: It should make all premises simultaneously true and hence show their consistency. But models can make much more than just the premises true unless they are partial models that do not specify a truth value for each input. J-L & B's notion of model is perhaps best characterized as a minimal model, that is, the intersection of all models that make the premises true, restricted to small finite domains and allowing for revision when more information is processed. Alternative models are represented on different lines (except when they illustrate spatial relations), as lines in truth tables for propositional logic represent different possible truth conditions of the atomic propositions. If the premises are about circles and triangles, their models are iconic. If the premises are quantified, they use *a*, *b*, and *c* to represent the structure. Universal quantifiers are represented by lists of [*a*] *b* and existential ones by similar lists of simply *a b*. The difference in meaning is indicated only by stating that the first representation is exhaustive and the latter not. The premise of a conditional is also represented by a bracketed symbol, though the connection to universal quantification is not made explicit. No rules tell us how to represent the content of the linguistic input in such models; only illustrations are given. Neither is the notion of inference characterized

in such a way that it applies to the different examples of models presented.

Surely logic gave us much better ways to represent the content of quantificational premises than this clumsy representational system! In fact, Chapter 9 in J-L & B's reasoning program does use variable binding and recursive procedures like substitution. It resembles a Prolog-style Horn clause logic program with negation. Now valid inferences are simply characterized as those arguments that have a true conclusion in every possible model where the premises are true and its semantics need not be limited to small, finite domains. If classical logical systems of representation fail to capture informative content expressed in a context, we should design new inference systems that do, using the heritage of logical research in the twentieth century, instead of starting from scratch.

***3. Rationality.*** "Rationality is problematical if it is supposed to be founded on rules. . . . The common denominator of rationality is the search for counterexamples. . . " (p. 209). Searching for counterexamples is not just a random trial and error process but rule-governed behavior. J-L & B's averseness to rules ignores the fact that they too rely, albeit implicitly, on rules that determine what constitutes a counterexample and how to find one. Rationality includes the cognitive capacity to conduct a systematic, rule-based search for counterexamples and to recognize one as such when it is found.

## Do mental models provide an adequate account of syllogistic reasoning performance?

Stephen E. Newstead

*Department of Psychology, University of Plymouth, Plymouth PL4 8AA, England*

**Electronic mail:** *p02111@pa.plym.ac.uk*

Mental models were first put forward by Johnson-Laird (1983) as an explanation of performance on syllogistic reasoning. Although the approach has since been extended to many other areas of cognition, it is still more fully articulated in this area than elsewhere. In this commentary I wish to examine how satisfactory the approach is as an explanation of syllogistic reasoning.

The main prediction made by the mental-model theory is that multiple-model syllogisms should be more difficult than single-model ones. Johnson-Laird & Byrne (J-L & B) have no doubt that this prediction has been confirmed, commenting in their accompanying Précis that "We have yet to test an individual who does not conform to this prediction" (sect. 6). This is a strong claim, but one which seems to be borne out by the evidence: A large number of studies have confirmed that the syllogisms J-L & B characterise as multiple-model problems are consistently harder than single-model ones. But there is a hidden problem here, concerning the way problems are characterised as multiple- or single-model syllogisms. A listing is presented in Johnson-Laird and Bara (1984), and is partly reproduced in *Deduction* (pp. 107–10). It is not clear to me, however, just how these models are constructed or, indeed, what the criteria are for models being distinct rather than variants of the same basic model.

Let me illustrate this with an example. The syllogism:

All A are B
All B are C

yields just the following model according to J-L & B:

[[a] b] c
[[a] b] c
. . .

This means that A is exhaustively represented with respect to B, B is exhaustively represented with respect to C, and there are other individuals not explicitly represented in the model (indicated by the three dots). From this single model representation, it is relatively easy to draw the conclusion that All A are C.

Consider what would happen, however, if subjects were assessing the validity of the conclusion All C are A. This is in fact not a valid conclusion, since the above model can be fleshed out as follows:

[[a] b] c
[[a] b] c
c

According to J-L & B, this is simply a fully fleshed out version of the same model; but it could be argued that this expanded version is actually a conceptually distinct model. It certainly leads to different permitted conclusions.

To illustrate the dilemma, consider also the syllogism:

All B are A
All B are C

This produces three models, as follows:

| (1) | (2) | (3) |
|---|---|---|
| [a [b] c] | [a [b] c] | [a [b] c] |
| [a [b] c] | [a [b] c] | [a [b] c] |
| | a | c |

It can be seen that the basic model is similar for all three models, the only difference being that in (2) the possibility that there might be A's which are neither C's nor B's is explicitly represented, and in (3) the possibility of there being C's which are neither A's nor B's is explicitly represented. What are the grounds for saying that these fleshed out models are conceptually distinct, whereas the fleshing out of the single-model syllogism discussed above is not?

It is of course perfectly possible that the distinction can be made, and indeed J-L & B indicate some of the ground rules that will need to be incorporated into a computer program which can construct mental models (pp. 177–80). This is a welcome development, but until such a program is developed there will be a lingering suspicion that the number of models a syllogism permits is to some extent determined by how difficult it is found to be.

Another area of syllogistic reasoning to which the mental-model approach has been applied is that of belief bias. In their Précis J-L & B are right in pointing out that any effects of believability are difficult to explain if reasoning is assumed to be based on formal rules. Their own research has confirmed the effects of believability and has also tested a specific prediction of the mental-model account: that belief bias effects should be stronger on multiple-model syllogisms than on single-model ones. Contrary to their expectations, belief bias effects were consistently found with single-model syllogisms. They did, however, find that the believability of conclusions deriving from the initial model constructed in multiple-model syllogisms affected the likelihood of subjects continuing to search for alternative, falsifying models; this is consistent with the predictions of mental-model theory.

Research in our own laboratory has extended this finding to explain the observed interaction between logic and belief in syllogistic reasoning. It has been found that the effects of logical validity are strongest on unbelievable problems (Evans et al. 1983). Mental models can explain this if it assumed that subjects seek to falsify a conclusion by searching for alternative models only if the first model they construct leads to an unbelievable conclusion. With single-model syllogisms, there is no alternative model to search for and hence the interaction should occur only with multiple-model syllogisms. A series of experiments

confirmed this prediction (Newstead et al., in press), though it should be borne in mind that the number of models underlying each syllogism was derived from Johnson-Laird and Bara (1984) and is hence subject to the criticisms mentioned above.

Although mental models provide the best account of the interaction between logic and belief, there are other aspects of the belief bias literature that they explain less well. For example, as mentioned above, belief bias effects occur with single-model syllogisms, contrary to the predictions of the theory. It has proved necessary to postulate the existence of a conclusion-filtering mechanism that is applied to all conclusions, regardless of the number of models underlying the syllogism (Oakhill et al. 1989). This is something of a face-saving exercise. There is nothing in the mental-model account itself that would predict such a mechanism, and indeed it has rather more in common with the response bias approaches that mental-model theorists have traditionally opposed.

There is in fact a way in which the existence of belief bias effects on single-model syllogisms can be explained that preserves the main aspects of the mental-model approach. One assumes that subjects construct their mental model of a single-model syllogism and then produce a conclusion consistent with this model. If the conclusion they produce is unbelievable, they may initiate a search for alternative models. They will not be able to find a genuine alternative model, but this process of search may be error prone, and on occasions subjects may believe they have found an alternative model and hence reject the (valid) conclusion they had originally produced. Since this search will not even be initiated if the initial model produces a believable conclusion, this can readily explain the observed belief bias effects.

The above comments are largely critical, but should not be taken to imply a general dissatisfaction with the mental-model approach. On the contrary, it is currently one of the most promising approaches to cognition; it is precisely because of this that it is important to explore its shortcomings and limitations.

## Mental models and the tractability of everyday reasoning

Mike Oaksford

*Cognitive Neurocomputation Unit, University of Wales at Bangor, Gwynedd LL57 2DG, Wales*

**Electronic mail:** *pss027@vaxa.bangor.ac.uk or mike@cogsci.ed.ac.uk*

In their new book *Deduction* Johnson-Laird & Byrne (J-L & B) present the results of their work extending the mental-model framework (Johnson-Laird 1983) beyond syllogistic reasoning to account for (i) other modes of deduction (Chs. 3–8) and (ii) everyday reasoning (Ch. 9). They argue (Ch. 9) that a mental-model account of (i) will smoothly generalise to account for (ii). Indeed, they view the processes of mental-model construction required to account for deductive reasoning (i) as already embodying mechanisms which account for everyday reasoning (ii). In this commentary I argue that mental-model theory is unlikely to shed any light on the core problems of everyday reasoning and that insofar as the mental-model account of deductive reasoning relies on a solution to these problems, that account is similarly suspect.

Everyday reasoning contrasts with deductive reasoning in being *nonmonotonic* or defeasible, that is, premises can be added and conclusions lost, or defeated. So, for example, when I turn the ignition key in my car I infer that the engine will start but this inference may be defeated if the battery is flat, the ignition is faulty, and so on. In monotonic, deductive reasoning, in contrast, the addition of premises cannot invalidate a previously valid argument. Most inferential performance which underpins human cognition is nonmonotonic (Oaksford & Chater 1991; 1992a). Thus, scientific inference to the best explanation (Fodor 1983), default reasoning (Reiter 1980; 1985), induction, abduction, eduction, and so on are all nonmonotonic. Problems for formal theories of nonmonotonic reasoning have emerged in the philosophy of science (Glymour 1987; Goodman 1983/1954; Quine 1953) and more recently in AI knowledge representation (McDermott 1986; Oaksford & Chater 1991). These theories have confronted two problems: (i) They typically fail to capture the intuitively correct inferences (McDermott 1986; Oaksford & Chater 1991), and (ii) these systems are computationally intractable (McDermott 1986; Oaksford & Chater 1991). Nick Chater and I have argued elsewhere (Chater & Oaksford 1993; Oaksford & Chater 1992a; 1992b) that mental-model theory fails to address either of these problems. Here I concentrate on the problem of intractability.

How intractability arises for nonmonotonic reasoning can be illustrated using an instance of reasoning to the best explanation (Fodor 1983). If my car doesn't start when I turn the key, then I am more likely to infer that the ignition is faulty than that someone has removed the engine overnight. The plausibility of the former conclusion over the latter, however, is only guaranteed relative to everything else I know (Fodor 1983; Quine 1953). Thus, if I knew that the ignition had just been changed and that there was a group of engine bandits operating locally, then the plausibilities of these conclusions may reverse and hence the default conclusion I should draw will change. Briefly, any default inference depends on its consistency with everything else that is known; consistency checking is an NP-hard problem (Garey & Johnson 1979); given that the whole of world knowledge may be implicated in any default inference (Fodor 1983), intractability will bite even for default inferences involving a single rule.

Mental-model theory does not seem to get off to a good start in resolving the problem of intractability, exemplifying as it does a strategy not noted for its success in AI. AI researchers have typically examined inference regimes in toy domains, involving small knowledge bases. In such domains these inference regimes may be shown to have some plausibility. However, they typically fail to scale up to more realistic settings involving inferences over large amounts of knowledge such as those implicated in everyday reasoning. Yet this is exactly the strategy that mental-model theorists propose. A theory that appears adequate to explaining reasoning performance in constrained laboratory tasks is to be scaled up to account for real-world nonmonotonic inference. These problems in AI suggest that the adoption of a similar strategy by mental-model theorists is unlikely to succeed, and this seems to be borne out in looking at the specific proposals made by J-L & B.

The principle innovation behind mental-model theory – that it is based on semantic rather than syntactic principles – exacerbates rather than alleviates the problem of intractability. Searching exhaustively for a countermodel to a particular argument is a less tractable procedure than its syntactic alternative even in the case of monotonic reasoning. Mental-model theorists are aware of this problem (e.g., Johnson-Laird 1983) and propose that mental models do not represent all possible models but instead only use *arbitrary exemplars* of the domains that figure in an argument. However, J-L & B, perhaps rightly (Oaksford & Chater 1992a), lay little emphasis on the use of arbitrary exemplars in avoiding the intractability of everyday reasoning, making an appeal instead to the ability to construct the right kind of mental model. Such an appeal, however, turns out to be circular.

J-L & B suggest that nonmonotonic reasoning is a natural byproduct of constructing mental models. Default assumptions, which can be undone in the light of subsequent evidence, can be recruited from prior world knowledge and embodied in a mental model. The problem of intractability can be avoided because no search for counterexamples to these default assumptions need be initiated and only a single representative model need be considered.

It is difficult to know how these proposals resolve the problem of intractability because no example of any benchmark defeasible inference (e.g., the Yale shooting problem, Hanks & McDermott 1985; 1986) is worked through in the book. (Indeed one suspects that the only reason J-L & B's proposals seem at all plausible is because they are worked through with an abstract example of spatial reasoning [pp. 181–83], which is unlikely to contact appropriate world knowledge.) So let us consider whether these proposals could resolve the defeasible inference we introduced above. The inference we want is that if the key is turned and the car does not start then I infer that the ignition is faulty rather than that the engine has been removed. Clearly a model in which the engine has been removed is less plausible than one in which the ignition is faulty. However, in J-L & B there are no inferential principles described which would indicate why the former rather than the latter model is constructed. It would appear therefore that the only grounds for differentiating these models is that a prior exhaustive search has been conducted which indicates that the situation where the ignition is faulty is more consistent with prior knowledge than a situation in which the engine has been removed overnight. But if constructing the right model involves such exhaustive memory searches, then mental models, per se, do nothing to resolve the problem of the intractability of everyday nonmonotonic inference. Moreover, since constructing the right model is itself a problem in nonmonotonic everyday inference, this "solution" is circular (Garnham [1993] makes a stronger case for the applicability of mental-model theory to nonmonotonic reasoning, but see Chater & Oaksford [1993]).

The situation is in fact worse: The mental-model account of deductive reasoning also relies on processes which construct just the right kinds of model. In J-L & B the explanation of the empirical data on deductive reasoning depends on the way an initial mental model of the premises is "fleshed out." "Fleshing out," for example, determines whether a disjunction is interpreted as exclusive or inclusive (p. 45); whether a conditional is interpreted as material implication or equivalence (pp. 48–50), which in turn determines whether inferences by *modus tollens* will be performed; whether nonstandard interpretations of the conditional are adopted (p. 67), including content effects whereby the relation between antecedent and consequent affects the interpretation (pp. 72–73); it also determines confirmation bias in Wason's selection task (p. 80) and the search for counterexamples in syllogistic reasoning (p. 119). The processes of "fleshing out" involve nonmonotonic reasoning, which, as J-L & B (p. 181) argue, "is essential for the program's operation [i.e., the program which constructs and manipulates mental models]; it is a process that is complementary to valid deduction." Thus, it would seem that rather than smoothly generalising to account for everyday reasoning, the mental-model theory of deductive reasoning presupposes a solution to the problems of nonmonotonic inference.

## Deduction and degrees of belief

David Over

*School of Social and International Studies, The University of Sunderland, Sunderland SR1 3SD, England*
**Electronic mail:** *os0dov@spock.sunderland.ac.uk*

Johnson-Laird & Byrne (J-L &B) have many interesting and stimulating things to say about deduction in their book. They are well aware that their account needs to be extended in a number of ways. I wish to focus here on what I see as one serious limitation of that account, hoping that they will agree that overcoming it is a priority.

J-L & B investigate what deductions people make from propositions they are *given* as premises. That is, people are presented with certain propositions, expected to assume that these are true, and required to answer certain questions about what validly follows. This may be done quite explicitly, as when they are given the premises of a possible syllogism and asked what follows. It may also be done rather implicitly, as when they are told to assume certain facts about four cards and then asked which ones need to be turned to find out whether a conditional is true or false. Some of these premises are abstract, arbitrary, and artificial; others are more realistic. But what subjects are basically being asked to do is itself fairly artificial. People generally use deduction in ordinary affairs to extend or deepen their beliefs for specific purposes. They do not often assume what they do not believe. They do admittedly do this sometimes. Ordinary people do occasionally use the form of *reductio ad absurdum*, which requires them to assume propositions believed by their opponents in an argument, with the intention of showing that an absurdity can be deduced from these. In decision making, people will assume that they have chosen one course of action rather than another, with the object of inferring what would follow from it. But this is a very special case, in that they can make such propositions true by actually choosing that course of action (after inferring that it is the best one for what they want).

Sometimes it is even pointless to assume that our actual beliefs are true, that is, to take them as certain, for the purpose of deducing interesting conclusions. To take a specific example, I now have a moderately confident belief about who will win the next American presidential election. *If* I assume that this belief is correct, and assume other general beliefs I have are true, I can deduce that the dollar will get stronger after the election. But do I now firmly believe that this will happen, and so go out and buy dollars? I do not, for I have made a number of assumptions for this deduction, and these might be false. It is true that these assumptions express my beliefs, but then I am not certain of these beliefs. I have more confidence in some than in others, and when I ask how likely they are to be all true together, I really have my doubts. There is nothing irrational in this; my degrees of belief could be coherent, in the sense of conforming to the principles of the probability calculus. But then it looks as though my deduction about the dollar was a waste of time, and indeed I would not have bothered to make it except for the purposes of an example.

In order to investigate deduction more realistically, we must have the means first of all to distinguish between assumptions and beliefs, and then to distinguish between, and theorize about, different degrees of belief. The representations of mental models in J-L & B's book could correspond to a person's assumptions or certain beliefs. A number of such representations could at best correspond to possibilities the person considers equally likely. They have no way as yet of representing mental models embodying more or less firmly held beliefs, along with other mental models standing for alternatives held to be less likely. J-L & B's theory does not yet cover the deductions of people whose mental representations are distinguished in this way. This is a serious limitation; just how serious can be illustrated by what they say about conditionals.

An indicative conditional tends to be asserted and accepted when its consequent seems highly probable given its antecedent. We need a way of accounting for this important point about these conditionals, which J-L & B do not give us. They consider the extreme view that such conditionals do not have truth conditions, but only assertibility and acceptability conditions. They reject this view because "if conditionals have no truth conditions then they cannot be true or false, and so they cannot occur in valid deductions, which by definition are truth preserving" (p. 65). This remark does not actually do justice to that view, which can define validity, using the fact that the uncertainty (1 − the probability) of the conclusion in a valid argument cannot exceed the sum of the uncertainties of its premises. (See Adams, 1975, on how to make this a rigorous definition.) In any case, we

do not need to take the extreme view to see the importance of examining people's deductive behaviour for different degrees of uncertainty in premises (where, again, one's degree of uncertainty is 1 − one's degree of belief). If we do not work with these concepts and distinctions, we are likely to be led astray.

J-L & B, for example, claim that "people do not have a secure intuition that modus ponens applies equally to any content." (Précis, sect. 4, para. 3) They conclude this from experiments run by Byrne (1989), in which subjects are given two conditionals and the antecedent of the first as assumptions, and the subjects do not deduce the consequent of the first as the conclusion. J-L & B claim that the second conditional "blocks" the application of *modus ponens* to the first conditional. The obvious reply, with the proper distinctions in hand, is that the particular second conditional the subjects are given greatly increases the imagined degree of uncertainty in the first. The subjects have no doubts about *modus ponens* for these conditionals, but in real life they do not waste time drawing conclusions from grossly uncertain propositions, and this carries over to the experiment. (J-L & B discuss only indicative conditionals in this connection. But one can make the case that *modus ponens* should not always apply to deontic conditionals. On this, see Evans et al. 1992 and Manktelow & Over 1992b. For more on Byrne's experiments, see Byrne 1991; O'Brien, in press; Politzer & Braine 1991.)

## Mental models, more or less

Thad A. Polk

*Department of Psychology and School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213*
**Electronic mail:** *polk@cmu.edu*

*Deduction* represents the most complete and accurate treatment of deduction to date. Johnson-Laird & Byrne (J-L & B) have provided explanations for behavior on all the standard tasks used in studying deduction and, in most cases, their accounts are more accurate than any previous theories. Furthermore, their explanations are derived from one general framework – mental-model theory. Consequently, the individual microtheories do not stand on their own, but provide mutual support for each other. In short, J-L & B have provided the first unified theory of deduction – an extremely important contribution to the study of cognition. Nevertheless, I have two major concerns about the book. Put simply, the authors' explanations of empirical results depend on both more and less than the basic mental-model theory. Consequently, the theory is neither as powerful as the book suggests nor as parsimonious as it could be.

The mental-model theory assumes that deduction consists of three stages: comprehension (constructing an initial mental model), description (formulating a parsimonious description of the models that have been constructed), and validation (searching for alternative models in which the putative conclusion is false) (pp. 35–36). The reader is given the impression that one can derive all the empirical results from this one basic theory. For example, J-L & B claim to have "formulated the first comprehensive theory in psychology to explain all the main varieties of deduction" (p. x) and that the empirical results "corroborated *the* model theory's predictions about propositional, relational, quantificational, and meta-logical reasoning" (p. 215, emphasis added). But in reality, the explanations of behavior on individual tasks rely, almost without exception on additional assumptions that are not part of the core theory.

The most obvious example is the prediction that the difficulty of metalogical problems is based on "the number of clauses that it is necessary to use in order to solve the problem" (p. 160). This prediction is based on the specific strategies J-L & B assumed subjects were applying and it clearly cannot be derived from the basic theory. The authors acknowledge as much: "the prediction is almost independent of the processing theory that we have proposed, and is likely to be made by any sensible analysis of meta-logical problems" (p. 160). Numerous other assumptions that are not part of the basic theory also play a role in explaining behavior. In Chapter 3, "negative deductions should be harder than . . . affirmative deductions" (p. 55). In Chapter 4, "subjects consider only those cards that are explicitly represented in their models of the rule" (p. 79). In Chapter 5, reasoners are assumed to consider multiple models even if both support the same conclusion (p. 96), and this seems to *contradict* the basic theory's assumption that additional models are considered only if they falsify the putative conclusion (p. 35). In Chapter 6, "it is harder to form an initial model in these (symmetrical syllogism) figures" (p. 123). In Chapter 8, "the hypothesis that an assertion is true should be easier to process than the hypothesis that an assertion is false" (p. 161).

The point here is not to question these additional assumptions (most of which seem quite plausible) or to criticize the model theory, but simply to point out that the predictions depend on both. J-L & B have developed a set of microtheories all of which are based on the same general framework. The fact that they could do so provides strong support for mental models, but it is far different from deriving all the predictions from that basic theory.

Although my first concern only involved the claims made about the theory, my second involves the theory itself. Validation (the search for alternative models that falsify putative conclusions) is at the heart of the theory, yet this stage does not seem to have much predictive power: Most of the predictions in the book do not make reference to it, and even those that do can be derived more simply without it. Consider the predictions about propositional reasoning which are summarized in the conclusions to Chapters 3 and 4. Most of these are based on the number of explicit models assumed to be constructed by comprehension. For example, deductions based on exclusive disjunctions are predicted to be harder than deductions based on conditionals. The reason is that the disjunctions are assumed to require the construction of two explicit models whereas the conditionals are assumed to require only one. But note that this prediction makes no reference to validation – it depends only on how many explicit models are produced by the comprehension stage. The other predictions in Chapter 3 are also based on assumptions about comprehension. Similarly, the predictions in Chapter 4 are not based on validation but on the initial models produced by comprehension: "These [initial] models lead to a defective truth table, an inability to make a modus tollens deduction, and a lack of insight into the selection task. When the models are fleshed out with explicit information, . . . then judgements conform to a complete truth table, modus tollens is deduced, and an insightful choice in the selection task becomes feasible" (p. 85).

Some of the predictions about relational reasoning (Ch. 5), on the other hand, do make reference to validation. For example, problem III (p. 96) is predicted to be difficult because any conclusion drawn from the initial model can be falsified by constructing an alternative model. In keeping with this prediction, only 18% of subjects get tasks like problem III correct. But the only reason problem III is predicted to be hard is that it requires validation. So according to the mental-model theory itself, less than 20% of subjects successfully apply the validation stage to these types of problems. And a substantial proportion of even these subjects may not be doing so. If they simply noticed the ambiguity of the premises during the initial encoding then they might respond "no valid conclusion" (the correct response) without ever having constructed alternative models.

Next, consider reasoning with quantifiers (Ch. 6 & 7). Once again, according to mental-model theory itself, most of the subjects do not successfully validate their conclusions. In the case of syllogisms, only 25% of responses are correct on prob-

lems that are assumed to require validation (p. 123). Similarly, in two experiments using multiple quantifiers, only 13% and 16% of responses were correct for valid multiple-model problems (pp. 139–40). Furthermore, Polk and Newell (1992) have constructed a theory of syllogisms that uses mental models and that can be parametrized to model individual subjects. Of the 103 subjects they studied, only 1 was fit significantly better with a validation strategy than without one. So even behavior that appears to be consistent with validation can be explained without it. J-L & B also point to subjects' memory of their responses as evidence for the search for alternative models (pp. 126–27). Subjects who responded "no valid conclusion" often claim to have given the response that would be predicted for a single model, presumably because they considered that model initially but rejected it during validation. But as the authors themselves admit, the possibility that these subjects "were reasoning from the premises once again . . . cannot be eliminated" (p. 127).

Finally, as I pointed out above, the predictions about metadeduction (Ch. 8) mainly depend on the metalogical strategies that subjects are assumed to apply. In any case, they do not depend on validating putative conclusions by searching for alternative models.

The reader may wonder why the issue of validation is important. After all, the mental-model theory is consistent with the idea that some subjects do not validate (because of working memory limitations). But recall that J-L & B consider validation to be at the very heart of deduction: "Reasoning itself depends solely on searching for counterexamples" (p. 144). So if, as I have argued, validation is very rare, then *Deduction* presents a distorted view of reasoning. An alternative view, proposed by Polk and Newell (1992), is that behavior on these tasks can best be characterized as *verbal reasoning*. They argue that encoding and reencoding are the central processes in the deduction of untrained subjects – not the search for alternative models.

Despite these concerns, *Deduction* represents one of the most important contributions to the study of reasoning in a long time. It presents the first unified theory of deduction and provides compelling evidence that rule-based theories are fundamentally wrong. In my opinion, Johnson-Laird & Byrne have laid to rest the issue of whether people reason using formal rules, context-specific rules, or mental models – they use mental models, more or less.

ACKNOWLEDGMENTS

Thanks to Richard Lewis for helpful remarks on this commentary.

## There is no need for (even fully fleshed out) mental models to map onto formal logic

Paul Pollard

*Department of Psychology, University of Central Lancashire, Preston PR1 2TQ, England*

Johnson-Laird & Byrne (J-L & B) provide extensive and convincing evidence that people derive deductions via mental models and do not possess formal inference rules. Personally, I am convinced that this is the case. However, although no formal logical rules are necessary for combining models or determining possible conclusions, I feel that there is an inherent inconsistency in the authors' position in that they appear to draw upon formal rules of inference in their theories of how models are constructed. This process is said to rest upon grammatical and semantic knowledge that allows an understanding of the statements to be modeled. For relational terms, it is clear how the meaning of terms such as "taller," "in front of," and so on, can be modeled from a clear understanding of the everyday use of such terms. However, in most reasoning situations covered in the book the statements to be modeled are somewhat more complicated. In every case, J-L & B assume that the model constructed (once fully "fleshed out") interprets the premises in the same way as they would be interpreted from a formal logical point of view. The instances of this that I will address here are models for "some" and "if . . . then."

Perhaps the most notable case in which subjects are assumed to build models based on formal logic is "if . . . then." Some of the most powerful arguments against rule theories arise from the discussion of conditionals, where it is argued that subjects do not even have a formal rule for *modus ponens*. However, paradoxically, this raises a problem for J-L & B's theory of conditionals as, if subjects have no rule for *modus ponens*, they should have difficulty constructing a model for "if then" and there is no reason why this model should accord with material implication. In particular, if *modus ponens* can be suppressed, there is no reason why a fully fleshed out model for "if $p$ then $q$" should not include some instances of "$p$ and not $q$." While arguing that subjects have no formal logical rules, J-L & B still propose a full model (i.e., once the implicit part has been fleshed out) of "if then" that accords with material implication. The effect of thematic content on the selection task is then explained as due to the content eliciting this full representation so that the impossibility of $p$ and not $q$ is apparent. However, thematic contents that have proved most facilitative on the selection task are all of a type that rely on natural rules about permitted or required activities.

Regardless of whether one wishes to explain performance in terms of a permission schema, it seems to me that these contents have one particular thing in common: In natural language they express conditional relationships that are true but for which counterexamples are known to exist. Thus, for example, it is true that there is a legal drinking age but it is also true that people drink when underage. If people don't have formal rules of inference then it follows that constructed models do not need to be based on them. Any full model of this rule should thus include, $p$ not $q$ cases, not as impossibilities, but as real tokens that are as likely to exist as $p$, $q$ instances. It seems likely that highly facilitative contents on the selection task work as they do because the mental model of the conditional includes explicit tokens for potential violations. It could be objected that any such model would render the conditional always true and thus testing it is pointless. However, the whole point about social rules such as the Griggs and Cox (1982) "drinks" problem and D'Andrade's Sears problem (Rumelhart 1979; 1980), and even the artificial problems created by Cosmides (1989), is that their truth or falsity cannot be determined by an evaluation of instances – as it is known that people break them. Facilitation is gained by asking subjects to search for violations and there is clear evidence (e.g., Jackson & Griggs 1990) that this focus on violations is an important component of the facilitation. In essence, I am suggesting that highly facilitative contents give rise to a model of the conditional that contains explicit $p$, not $q$ tokens, thus allowing these to be easily identified by solvers when they are asked to search for violations. Allowing "nonlogical" models of this type adds explanatory power and can be viewed as a necessary consequence of the argument that people do not have formal inference rules.

Throughout discussion of syllogistic and multiply quantified arguments, the description of a fully explicit model of an existential premise relies on its interpretation in formal logic as "some and possibly all." However, most subjects appear to interpret "some" as meaning "some are and some are not," even when explicitly instructed as to its meaning in formal logic (e.g., Newstead 1989). It follows clearly from J-L & B's overall theory that subjects should construct models using this everyday interpretation. In particular, many subjects will not construct models that include the "all" possibility. Thus, subjects on some multiple model problems will fail to generate all the possible models not because of processing limitations, restricted search, or whatever, but because for them these models simply do not

exist. Failure to construct certain alternative models is then another way of saying that the subject's interpretation of the premise is discrepant with its specification in formal logic. Greene (1992) has presented a reinterpretation of the results from the problems using doubly quantified premises that is based on an analysis of the subjects' interpretation of the premises used. In their reply, one argument that Johnson-Laird et al. (1992) use is that this reanalysis fails to explain why interpretational difficulties are confounded with the number of models needed. However, the reliance on the logical rather than the natural interpretation of quantifiers on the proposed model construction is in danger of leading to circularity. Does the number of models needed account for apparent interpretational differences, or do interpretational effects account for apparent differences between models? "Misinterpretations" (from the perspective of formal logic) are bound to lead to conclusions that are consistent with some, but not all, models (a finding J-L & B frequently cite). Thus, misinterpretation of some premises is not inconsistent with the results and could be incorporated within the theory.

It seems to me that in restricting models to the logical interpretation, J-L & B have introduced an essential contradiction in the overall theory and missed the opportunity to enhance its explanatory power, both in relation to the selection task and in relation to the incorporation of natural language interpretations of premises in reasoning with quantifiers.

## Unjustified presuppositions of competence

Leah Savion

*Department of Philosophy, Indiana University, Bloomington, IN 47405*
**Electronic mail:** *lsavion@iubacs.bitnet*

The theory of mental models (MM) provides a partial account of the effect of the semantic interpretation of premises on the conclusion drawn and brings to light the role of imagery in reasoning. The theory is seriously incomplete on several scores, notably at the "algorithmic level." For example, there is no way of accounting, within the theory's framework, for immediate "automatic" inferences people generally make from formally stated premises.[1] Also, explaining inferential failures in terms of incomplete representation of the premises leaves little room for the many cognitive heuristics and conspicuous human biases involved in reasoning.

Mental-model theory, however, cannot be saved by patching up these and other incomplete accounts. In my opinion, the theory is intrinsically inadequate as a model of actual human deductive reasoning. In attempting to replace the unreasonable assumptions of mental logic theories by postulating "psychologically real" assumptions, Johnson-Laird & Byrne (J-L & B) end up presupposing an enormous core of competence that contains declarative knowledge, procedural knowledge and skills, all under the rubric of "rationality." These implicit components are necessary for each stage of the deduction process: construction of models, inference, and search for counterexamples. Most of these assumptions do not stand up to intuitive judgments or to simple empirical observations. The following make some of the major assumptions explicit.

A. Competence consists of the central "core of rationality, which appears to be common to all human societies. It is the semantic principle of validity: an argument is valid only if there is no way in which its premises could be true and its conclusion false" (p. 209). As a matter of fact, the principle of validity is one of the most difficult concepts to teach to laymen. In elementary logic courses, I have observed thousands of college students demonstrating a deep-rooted confusion between the criterion of deductive correctness (validity) of reasoning and the truth of the sentences involved. This misconception leads many to the conviction that no correct conclusion is (factually) false, and further – that no correct reasoning can be exercised on false premises.[2]

B. The ability to search for counterexamples for the initial conclusion drawn is postulated by the MM theory as an essential ingredient of our competence core that is carried out at the last stage of deduction. This "universal" inferential strategy presupposes the understanding of the notion of validity and its subtle connection with the notion of truth. The naive reasoner is not normally aware of the instruction that the definition of validity implies: Always assume all the premises true in order to see whether the conclusion follows on the basis of that assumption. The search for a counterexample, we are told, proceeds by attempting to construct models that falsify the putative conclusion. But if the conjunction of the premises with the denial of the conclusion yields a contradiction, the untutored person may conclude that one of the premises is false. Furthermore, the application of this strategy requires a clear conception of the notion of *contradiction*. Alas, most naive reasoners offer the sentence "no woman is smart" as contradicting the sentence "all women are smart," demonstrating a confusion between contradiction and incompatibility.

C. The MM explanation of reasoning with quantified premises makes several assumptions about people's competence that prove to have no psychological reality. For example:

(1) People distinguish the instantiation of a general sentence (into an arbitrary object) from the instantiation of a particular sentence (into a nonarbitrary object).

(2) People are always clear about the contextual meaning of the ambiguous copula ("is," "are") that can be used to indicate inclusion (as in "all cats are animals") or the stronger relation of identity ("all lawyers are attorneys").[3]

(3) People do not construct unwarranted additional models. For example, the model for "some students did not pass the test" should not contain the information "some students passed the test." In reality, many fallacious inferences are drawn not because of incomplete representation of the premises, as J-L & B claim, but rather from misconception of the semantic-logical properties of quantity indicators that trigger "implicature" understanding and enlargement of the information given.

D. The MM account of propositional reasoning makes similar unrealistic assumptions about the elementary universal logico-semantic competence that people bring to bear in their reasoning. Building the suggested internal models of conditional premises (whether explicit or implicit) presupposes the possession of the notions of sufficient and necessary conditions, which, unfortunately, most lay people lack. In general, we are told that all that people need for the construction of models is "a knowledge of the meaning of the connectives and other logical terms that occur in the premises" (p. 39). Assuming that we all know the meaning of logical terms would blatantly beg the question. The solution offered – that these logical aspects of meaning "emerge" from the use we make of these terms in constructing and interpreting the relevant models (pp. 92, 103, 145) – takes us straight into the fire.

J-L & B suggest a refreshing departure from the idealistic syntactic approach to human reasoning. Unfortunately, the MM theory's model of actual reasoning is highly idealistic. If the acquisition of deductive competence is "profoundly puzzling" for formal rule theories (p. 204), it is astounding for the MM theory. By presupposing the above components of competence and others,[4] the MM theory makes intolerable demands on human competence, thereby failing to deliver its promise to "show how a theory based on models is able to account for deductive competence and for systematic patterns of performance" (p. 131).

NOTES

1. Hardly anyone fails to conclude "all *A* are *C*" from "all *A* are *B* and all *B* are *C*." The improvement of one's logical skills depends heavily on skills that manifest themselves in such "automatic" responses.

**2.** One of the major arguments proposed against the rule theory is the phenomenon mistakenly labeled as "rule suppression" (p. 199): Given that if Lisa goes fishing she'll have a fish supper, and that Lisa goes fishing, people are reluctant to conclude that she'll have fish for dinner once they are told that Lisa may not catch any fish. The story does not prove the nonuniversality of the "suppressed" *modus ponens*. Rather, since the first premise of the argument was rendered false by the additional premise, people simply refuse to use it in reasoning.

**3.** Evidence indicates that when given a neutral or arbitrary-content premise, people make a completely arbitrary choice between the two meanings.

**4.** For example, the layman is supposed to possess the semantic principle of compositionality, the deduction theorem and the working of recursive functions.

## Nonsentential representation and nonformality

Keith Stenning and Jon Oberlander

*Human Communication Research Centre, University of Edinburgh, Edinburgh EH8 9LW, Scotland*

**Electronic mail:** *k.stenning@ed.ac.uk; j.oberlander@ed.ac.uk*

The following appears near the end of Chapter 1 of *Deduction* and encapsulates Johnson-Laird & Byrne's (J-L & B's) position:

> No practical procedure can examine infinitely many models in searching for a possible counterexample to a conclusion. Hence, what logicians have proposed are systems of *formal rules* based on the idea of such a search for counterexamples. . . . But, the rules operate at one remove from models: they manipulate logical forms as do the rules of a natural deduction system. What we aim to show in this monograph is: 1. that in everyday reasoning the search for counterexamples can be conducted directly by constructing alternative models. (p. 16)

It might be thought that the primary claim is that the *sentential* nature of the representations logicians use is inessential to their craft. But then the fact that one can devise nonsentential reasoning systems – such as the well-known graphical logic diagrams of Euler, Peirce, and Venn – would mean that J-L & B's book was uncontroversial. One might perhaps then argue that mental models are not as nonsentential as they seem; there are simple algorithms for presenting them in a more sentential format. But this would be a minor quibble; the claim that sentential representations are not privileged has been made a number of times, most notably and with considerable rigour by Barwise and Etchemendy (1991).

However, this clear claim does not appear to be the object of J-L & B's book. Consider the occurrence of "directly" in the final sentence of the quote. The authors apparently believe that their reasoning system does not involve representations at all, or that what representations it does use are not governed by rules of manipulation and interpretation. Hence, they appear to believe not only that their system is nonsentential, but also that it is nonformal.

Now, the history of Johnson-Laird's work suggests the origins of this mistake. That work has been driven by the important insight that much of human reasoning is not deductive and that even when the task is deductive, performance varies substantially with the content of problems. It is therefore an important goal to explain how dependence on content arises. J-L & B maintain that content-dependence cannot be explained without the concept of nonformal reasoning, and that mental-model theory (MMT) is nonformal in just the sense required. The book is part of an attempt to show that MMT can explain aspects of content-dependence. However, we would argue that (i) MMT is actually formal, and that (ii) it fails to account for content-dependence. If it fails, however, its formality is not to blame, since (iii) one can construct an explanation of content-dependence compatible with formality. Leaving point (ii) to one side, let us consider the other points in turn.

On point (i), we would say that MMT is purely formal, though nonsentential, since it essentially provides unconventional proof-theories for various logical fragments. The theory does involve representations that are manipulated formally, and like all proof-theories, these manipulations are designed to allow computational access to an underlying model theory. The real contribution is not that these procedures "manipulate models directly," but that mental-model notation captures the fact that human strategies of reasoning are highly *agglomerative* (see Stenning 1992).

On point (iii), it is this insight concerning agglomeration that indicates where explanations of at least some forms of content-dependence must focus: on the nature of the underlying working-memory binding mechanisms. It is, of course, important to observe that other forms of content-dependence can be accounted for at a "higher" level: Reasoning with the syllogistic premises "All women are female" and "All men are women" would indeed be affected by taxonomic background knowledge. But in some cases, we can indeed trace the content-dependence of reasoning to the attribute binding mechanisms in working memory. It is these which implement the proof theory; and it can be shown that they give rise to its agglomerative style (see Stenning & Oaksford, in press; Stenning & Oberlander, submitted).

What follows from accepting points (i) and (iii)? from accepting that MMT gives us a formal reasoning system, and that explanations of content-dependence can be located at an implementational level? One conclusion is that mental models' main interest lies in their nonsentential nature, as we suggested to begin with. This should remind us that there are traditional graphical methods of syllogistic reasoning which are more thoroughly nonsentential than MMT. Where mental models could naturally be represented as 1-D list structures, the graphical methods exploit 2-D geometrical-topological relations. One such method – "Euler's Circles" – can be shown to be equivalent, when properly interpreted, to mental models (Stenning & Oberlander, submitted). Because the graphical representations are more constrained than mental-model notation, however, they also reveal novel features of human performance on syllogisms, allowing, for example, a generalisation of the "figural effect" (see Yule & Stenning 1992).

If J-L & B's aim was to champion nonsentential methods of reasoning, then it is curious that they go to considerable lengths to rule out graphical methods at a number of points. For example, they argue in Chapter 7 that since Euler's Circles treat monadic arguments, they cannot treat relational arguments. Yet the mental-model treatment they present uses page layout in a manner transparently adaptable to Euler's Circles. The truth is that the relational arguments the authors pick are trivially reducible to the Euler's Circles monadic fragment. And their choice of arguments is no accident; it is this precise fragment that MMT actually treats. One is left with the impression that J-L & B are simply trying to preserve the apparent originality of their system, even though it is arguably a somewhat less perspicuous version of a traditional graphical technique.

If, on the one hand, mental-model theory is taken as an argument for the liberation of psychological work on human reasoning from an outdated mechanical view of what logics are, then it has made a valuable contribution. As such, it points the way towards more radically nonsentential theories of human reasoning. If, on the other hand, mental-model theory is taken as an argument that human reasoning is computational but nonformal, dealing directly with the world without the intervention of representations, then it is downright misleading. As such, it needs to be further informed with concepts from the disciplines which deal with representation, syntax, semantics, and computation.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the support of the Joint Councils' Initiative in Cognitive Science and HCI through project number

G9018050 ("Signal"). HCRC is supported by the Economic and Social Research Council.

## Models, rules and expertise

Rosemary J. Stevenson

*Department of Psychology, Human Communication Research Centre, University of Durham, Durham DH1 3LE, England*
**Electronic mail:** *rosemary.stevenson@durham.ac.uk*

Given the extensive use of evidence and argument with which Johnson-Laird & Byrne (J-L & B) make their case for mental models, it seems to me to be hard to dispute the broad sweep of their position. Nevertheless, I will argue that mental models is not a truly general theory of the performance of individuals who have no formal training in logic. I will also suggest that even if it were such a theory, it would still be valuable to have an account of how untutored novices may turn into logical experts. Indeed, the beginnings of such an account of the acquisition of expertise may lie in J-L & B's own work on the metadeductive abilities of subjects untutored in logic.

In the context of deduction, I use the term "expertise" with two distinct referents: *substantive expertise*, which is expertise in the content domain being reasoned about, be it physics (e.g., Larkin et al. 1980) or racetrack handicapping (Ceci & Liker 1986); and *logical expertise*, which is expertise in logic itself. Substantive expertise is relevant to the argument that the theory of mental models is not a general theory of everyday deduction. Logical expertise is relevant to the suggestion that untutored novices may turn into experts in logic.

Expertise in a content domain may result in once difficult deductions becoming automatic in the manner described by Anderson (1982). In other words, people may acquire domain-specific production rules that automatically compute routine deductions in a highly familiar domain, thus leaving working-memory capacity free for developing new ideas through the use of mental models. Domain experts may therefore use mental models only when developing new ideas, instead using domain-specific production rules when reasoning with familiar content in their domain. Such experts need not be experts in logic. Thus their performance legitimately falls within the bounds of a theory of untutored deduction. J-L & B's reference to increases in the capacity of working memory during childhood (p. 204) also neglects the potential importance of knowledge. Chi (1978), among others, has argued that these capacity increases reflect increases in knowledge with increasing age rather than general increases in capacity. In ignoring substantive expertise, J-L & B may have neglected a major factor underlying deduction even in people untutored in logic. A proper test of the generality that is claimed for mental models therefore requires that the experimental subjects include substantive experts as well as novices in the domain specified by the content of the experimental materials. The routine use of neutral or abstract content in experimental materials (except when belief effects are under scrutiny) may effectively mask the potential effects of substantive expertise, as may the failure to distinguish subjects in terms of such expertise.

Expertise in logic may result from either explicit tuition or a protracted period of self-reflective metalogical processing of the kind discussed by J-L & B. With sufficient experience in solving logical problems, coupled with self-reflection, it may well be that seasoned reasoners can extract from their performance the formal rules that describe validity and subsequently use these rules instead of mental models. Such reasoning by rule may be what enables logical experts to focus on the form rather than the content of a problem when deriving a valid conclusion, an ability that is notoriously difficult for nonexpert reasoners. The beginnings of such a process may be evident in studies that show the rudimentary use of rules in solving syllogisms (e.g., Galotti et al. 1986) and relational problems (e.g., Wood 1969), although true expertise takes longer than the duration of a single experiment. True experts in logic may hence use formal rules for the kinds of abstract or neutral problems typically presented to experimental subjects, although they may need to use mental models to develop new ideas. Thus, a proper test of the use of formal rules of inference may call for subjects who have logical expertise as well as subjects who are untutored or otherwise inexperienced in logic.

J-L & B might justifiably argue that even if the arguments about logical expertise are plausible, they do not affect their theory because it was not intended to account for such expertise. Nevertheless, their account of metadeduction depends crucially on the ability to reflect on one's own and other people's deductions, and their account of development assumes that such abilities develop naturally (p. 204). Yet the development of metacognition is decidedly problematic (e.g., Brown & Day 1983). Abilities of the kind assumed by J-L & B in their account of metadeduction do not appear to be an automatic consequence of development. Indeed, attempts to train such skills also show limited success (e.g., Brown et al. 1981). As Bereiter and Scardamalia (1989) point out, the very existence of study skills training programs, typically aimed at college students, testifies to the fact that many students reach young adulthood with inadequately developed strategies for reflecting on and monitoring their own cognitive processes, other than those required for the memorization of facts. In other words, the subjects who participate in experiments on metadeduction may be ill-prepared for the task that confronts them. The identification of strategies that subjects use to solve metalogical problems (pp. 155–59) does indicate how novices approach a task in which they have little prior experience, and this of course was the intended scope of the work. But such strategies might have greater theoretical significance if they were located in a context that allowed an examination of how they might be used in the acquisition of logical expertise. For that, a characterization of the ways in which logical experts also solve such problems is needed.

## Scientific thinking and mental models

Ryan D. Tweney

*Department of Psychology, Bowling Green State University, Bowling Green, OH 43403*
**Electronic mail:** *tweney@andy.bgsu.edu.internet*

Simplicity and elegance are not all that common in cognitive science; because so much thought is "messy" – particularized, contextually dependent, multicomponential – our theories of thinking have tended to share the same attributes. It is refreshing, therefore, to see an exception. Johnson-Laird & Byrne (J-L & B) have given us a simple and powerful account that bridges a great many separate phenomena in deduction, is pared down to its essentials, and provides an array of examples that simultaneously illustrate the use of their approach and establish its generality across a wide range.

Any theory of thought must, however, allow us to bridge the gap between the pared-down laboratory implementations that are the focus of most research and the messier real world of thought. Although J-L & B spend little time doing this in the book, it is not difficult to see that the account can be used in such extensions. In part because of its simplicity, the mental-model theory of deduction permits us to see how far it can go into the realm of ordinary and extraordinary real-world thinking.

My own concern has been with the nature of scientific thinking (Tweney 1989; Tweney et al. 1981). In this realm, the argument of J-L & B has easy applicability. Introspective ac-

counts of science are of course easy to handle. Virtually any scientific diary can be read as an articulation of a mental model; the semantic account of deduction (that deduction amounts to a search for counterexamples) is enshrined in methodological texts; few scientists ever study formal logic or talk as if they had an intuitive grasp of its principles. Rejections of formal logic as a model of thought are easy in this postmodern age in which logical positivism is almost universally rejected by scholars in the study of science (except as whipping boy). Alas, the alternative view is frequently one in which scientific thinking is dismissed as irrational, there being an easy dichotomy between two views: scientist as dispassionate logical engine versus scientist as wildly irrational, driven by self-interest and dreamy intuition. Reality is between these extremes; most scientific thought fits comfortably with the framework established by J-L & B, a recursive search for counterexamples to models and the consequent transformation of those models. Sometimes that is all scientific thinking is, as when thought experiments are used to settle a point or advance an argument.

If J-L & B have thus accommodated scientific deduction in their account, perhaps there is another aspect of scientific thinking that, by being less easily accommodated, might also reveal the incompleteness of the explanation. Science is not just the evaluation of models; first, the models must be built. Here, the interactive nature of thought comes to the fore: Models emerge dynamically from the the to-and-fro between scientist and nature, a process of negotiation. If J-L & B's book helps us to conceptualize one piece of this process, it does not help much in understanding the protracted and dynamic character of the negotiation as a whole. Their theory thus runs up against the same limits as that of its never-named but obvious progenitor, Sir Karl Popper; the mental-model theory of deduction works only with finished mental models and says nothing about the process of constructing those models. Popper dismissed such concerns as mere "psychology"; J-L & B relegate them to other areas of cognitive science. But scientific thinking is not so easily compartmentalized. Do we really believe that perceptual processes (say) are separate and prior to the manipulation of the mental models they evoke? Don't we really expect that perception is in part driven by the existence of such models? And, if so, how are we to incorporate such things in the mental-model theory? How, in short, can we accommodate the simple match between model and data without losing the very simplicity that constitutes part of the charm of this account?

The question is partly answerable by appealing to other kinds of theory. We might, for example, imagine something like a neural network that drives the perceptual process with something like backpropagation as feedback from the mental model. The search for counterexamples could then be "symbolic" in the manner of Newell and Simon, but mated to a "subsymbolic" interface with the real world. All this, however, shows the incompleteness of the explanation here given us. The theory is in danger of becoming messy.

My quibble therefore is a small one and back-handed in character. Elegant though the J-L & B theory is, will its elegance survive its inevitable extension to more complicated cases?

## More models just means more difficulty

N. E. Wetherick

*Department of Psychology, King's College, Old Aberdeen AB9 2UB, Scotland*

Johnson-Laird & Byrne (J-L & B) hold (rightly, in my opinion) that deduction is of central importance in cognition generally. They distinguish three possible types of explanatory theory: formal rules, content-specific rules, and mental models, showing that mental-model theory outperforms the other two. They go on to assert (wrongly, in my opinion) that these three types exhaust the possibilities (e.g., p. 194). No argument is offered in support of this assertion and in fact the three are merely the types of theory currently under investigation in the fashionable journals. For an alternative, see Wetherick (1989; in press). Even if the assertion were true, mental-model theory remains open to criticism at several levels.

We need to know, first of all, what exactly a mental model is. Are there any mental representations that are not models, and if so what are they? J-L & B suggest (p. ix) that mental models resemble the models constructed in perception – as if it were a self-evident truth that perception involves the construction of models (it is not). But if it were, this would suggest that the models are in consciousness. The authors, however, do not insist that mental models are in consciousness. If they did, the theory would be immediately refuted by the observation that syllogisms, for example, may be solved with nothing in consciousness resembling a mental model (Wetherick 1991). Mental models are, however, in working memory. Working memory capacity is in fact the principal constraint on efficiency in deductive problem solving, but this does not help us much, because the relationship between working memory and consciousness is also obscure. I can understand what it would be like to construct a mental model in consciousness and evaluate putative conclusions against it (although the theory ought then to consider the activities of the constructor/evaluator, and these activities J-L & B pass over in silence). I cannot attach any meaning to the idea of "constructing" and "evaluating" a mental model that is not in consciousness (i.e., one of which I am not consciously aware). It cannot in any case be claimed (p. 212) that mental-model theory "makes no use of modus ponens." No theory that applies general principles to particular cases can avoid it.

As J-L & B emphasise repeatedly, every experimental test they have conducted showed that problems requiring the construction of more mental models were more difficult to solve. From this it follows that we need an objective means of establishing how many models are required for a given problem. No such objective procedure is offered and it does not inspire confidence to find that, for example, the syllogism AAIf.3. ("all $m$ is $p$" and "all $m$ is $s$" so "some $s$ is $p$") was said to require two models in Johnson-Laird (1983), one in Johnson-Laird and Bara (1984), and three in the present volume. Three models locates it accurately among the "very difficult" syllogisms (mean rank 21.7 ex. 27).

J-L & B show that they are aware (p. 194) that an experimental finding supports any theory from which the finding may be predicted but they fail to consider obvious alternative explanations of their own results. One typical experiment (on syllogistic reasoning) contributes to Table 6.1 and was reported in detail in Johnson-Laird and Bara (1984); it is also the one on which the main argument rests in Johnson-Laird (1983) and one of those selected by Newell (1990; see also multiple book review, *BBS* 15(3) 1992) to demonstrate the power of SOAR to simulate representative experimental findings in cognitive psychology. In it, 20 subjects were asked to attempt all 64 moods of the syllogism (27 of which are valid). In 19 of the 27 valid moods, one of the premises is an A proposition ("all $s$ is $p$") and the other, either I ("some $s$ is $p$"), E ("no $s$ is $p$") or O ("some $s$ is not $p$"). There are altogether 28 syllogisms answering this description and in 14 of them the valid conclusion is of the same logical form as the second premise (A, I, E, O). Of the remaining 14, 5 have a valid conclusion of a different logical form from either premise (one I, four O) and 9 have no valid conclusion at all. We have here two sets of 14 syllogisms that are identical regarding form of premise. In one set the valid conclusion is always of the same form as one of the premises and in the other set this is never the case, thought it is of course still possible to propose such a conclusion as an incorrect response. That is what most of the subjects did. To a syllogisms presenting two A premises, 90% proposed an A conclusion when it was correct and 60% did so

when it was wrong. To an A and an I premise, 79% proposed I when it was correct and 73% did so when it was wrong. To an A and an E premise, 74% proposed E when it was correct and 75% did so when it was wrong. To an A and an O premise, 26% proposed an O when it was correct and 58% did so when it was wrong!

This seems to me to be conclusive evidence that in both sets these subjects were simply matching the logical form of one of the premises; they had not engaged at all with the logical task. (In my own studies [e.g., Wetherick & Gilhooly 1990] some subjects behave in this way but some engage with the logic; the crucial indicator is the degree of success achieved on syllogisms having a valid conclusion that is not of the same logical form as either premise.) J-L & B unhesitatingly accept their subjects' performance when it is correct and ignore it when it is wrong. The set of 14 syllogisms in which the valid conclusion is of the same logical form as one of the premises happens to comprise the 10 said to require one mental model and the 4 (with an A and an O premise) said to require two. Thirteen syllogisms are said to require three mental models; that is, the 5 already mentioned (to which no correct conclusions were proposed) and 8 having an I and an E premise and a valid O conclusion (to which 18 correct conclusions were proposed, ex. 160). They all have valid conclusions of a logical form different from that of either premise, conclusions that cannot be obtained by matching; hence the apparent relationship between problem difficulty and number of models required. Neither the correctness of correct responses nor the incorrectness of incorrect responses has been shown to have anything at all to do with number of models.

I conclude that mental-model theory has not been shown to have any claim on our attention as a scientific theory explaining deduction. "Requires more models" seems simply to be an alternative way of saying "more difficult" that contributes nothing to our understanding of the nature of the difficulty.

# Authors' Response

## Mental models or formal rules?

Philip N. Johnson-Laird[a] and Ruth M. J. Byrne[b]

[a]*Department of Psychology, Princeton University, Princeton, NJ 08544 and* [b]*Department of Psychology, Trinity College, University of Dublin, Dublin 2, Ireland*

**Electronic mail:** [a]*phil@clarity.princeton.edu;* [b]*rmbyrne@vax1.tcd.ie*

> I never read a book before reviewing it; it prejudices a man so.
> Reverend Sydney Smith (1771–1845)

Our argument is that deduction is a semantic process. Reasoners understand some information, formulate a conclusion, and test its validity. To understand is to construct mental models from knowledge and from perceptual or verbal evidence. To formulate a conclusion is to describe what is represented in the models. To test validity is to search for alternative models that refute the putative conclusion. This theory contravenes received wisdom. Current theories assume instead that deduction is a syntactic process and that the mind uses formal rules of inference to manipulate representations of the logical form of premises. One formal rule that exists in the mind, according to nearly everyone, is *modus ponens*:

If p then q
p
Therefore, q.

The mind may also use rules with specific contents. *Deduction* reports experiments in the main areas of deductive reasoning, and their results corroborate the model theory and count against the psychological theories based on formal rules.

A gratifying number of reviewers accept our argument, but some have misunderstood it, and so we will clear up their misapprehensions before we reply to our critics. Some reviewers take us to have argued (unsoundly) that the model theory is better than any other possible theory (**Braine, Fetzer, Wetherick**). In fact, our claim was that the theory is better than other existing psychological theories, not that it is God's truth. We did not even claim to have excluded all possible theories based on formal rules – after all, a computer program implementing the mental-model theory *is* a formal rule theory (as **Bundy, Inder,** and we ourselves point out, *Deduction* p. 213). And we certainly did not argue against rules per se (*pace* **Andrews, Stenning & Oberlander,** and **ter Meulen**). Comprehension and reasoning rely on syntactic rules, semantic rules, rules for constructing models, and so on. If readers feel happier referring to them as "tacit inference rules" (**Braine**), so be it.

To those who believe there may be a superior theory based on formal rules or some tertium quid (**Bach**), we can only say: we agree, but until someone formulates such a theory, the point is not a lot more thrilling (to us) than the claim that any scientific theory may be superseded. To those who believe that we argued for the model theory over all other possible theories including formal or syntactic ones, we can only cite the following passages: "No amount of data, of course, can pick out one theory against all comers" (p. 194); "Mental models and formal rules both depend on syntactic procedures. . . . This claim is true for the computer programs modelling both sorts of theory" (p. 213).

One other misunderstanding arose about the formal rule theories. They postulate that difficulty reflects the number of steps in a deduction, and the relative availability, or ease of use, of the relevant rules. Commentators take us to task over these measures: they do not reflect differences between direct and indirect proofs (**Fetzer**), and a better measure is the number of embedded assumptions (**Crawford**). The measures, however, were *not* of our devising, but proposed by Braine, Rips, and other rule theorists (see pp. 29–31), who will doubtless appreciate these criticisms. Even so, the new measures fail, for example, to save the formal theories of spatial inference.

In our account of formal rules, we concentrated on so-called natural deduction systems. We did not describe Hintikka's (1955) model-set method, or the related "tableau" methods of Beth (1955) and Smullyan (1968), though we did refer to them as formalizing the search for counterexamples (p. 16; *pace* **Fetzer**). The reason for our focus was that psychologists have not adopted tableau methods. **Andrews** regards mental-model building as a kind of tableau development; and **Grandy** emphasizes that disjunctions increase difficulty for both models and tableaux. There is a resemblance, but as Andrews points out there are differences – particularly in our use of implicit representations. Fetzer even constructs tableau proofs with lengths that do not predict difficulty – human reasoners are evidently not using these particular tableau rules!

So much for the main misunderstandings of the book. Our present goal is to help readers – and reviewers mindful of Reverend Smith's advice – to make up their minds about it. Its argument still stands, and, thanks to the commentators, it can be clarified and strengthened. We will concentrate on the controversial issues, but to those who agreed with us we are grateful for their encouraging words. We will try to deal in passing with those notions that people found difficult to understand because our exposition was unclear. We begin with deductive competence and the algorithmic theory of performance. We then discuss each of the separate areas of deduction: propositional, relational, quantified, and informal reasoning. Finally, we consider rationality, the theory's extensions and deficiencies, and the possibility of combining it with formal rules.

**R1. The nature of logical competence.** Have we overestimated the importance of logical ability? Some reviewers think so (**Fisher, Galotti & Komatsu,** and **Luchins & Luchins**). Unfortunately, apart from the validation of certain intelligence tests, there are no objective data. We did concede that other forms of thought, such as creativity, are more important, but they are also much harder to understand. We believe that a world without deduction would be a world without science, technology, laws, and social conventions (p. 3). And if the controversy about it is not resolvable, what hope is there for cognitive science?

In characterizing deductive competence, we argued that people are rational in principle, but they transcend logic because their conclusions are parsimonious, maintain semantic information, and establish propositions not explicitly asserted in the premises. If no conclusion meets these constraints, then people say nothing follows. Hence, contrary to **Fetzer**'s suggestion, we do not believe that formal systems are normative of human competence: they allow infinitely many different valid conclusions from any set of premises, including conclusions that people would never draw. Most commentators appear to accept our account of competence, but three reject it. Their objections do not seem to be decisive.

*Modus ponens* throws away semantic information, and so **Cohen** concludes that we are wrong, either to assume that it is a respectable deduction or to hold that deduction ought to maintain semantic information. In fact, we deal with *modus ponens* (p. 22):

> A special case of parsimony is not to draw a conclusion that asserts something that has just been asserted. Hence, given the premises:
>
> If James is at school then Agnes is at work.
> James is at school.
>
> the conclusion:
>
> James is at school and Agnes is at work.
>
> is valid, but violates this principle [of parsimony], because it repeats the categorical premise. This information can be taken for granted and, as **Grice** (1975) argued, there is no need to state the obvious.

It is therefore not a counterexample to our account: there is no need to reassert the categorical premise.

**Cohen** also claims that our studies presuppose rational competence, and that our argument would collapse if this assumption had to be abandoned. We agree that one cannot use errors to corroborate a theory without knowing what counts as an error. We presuppose the semantic principle of validity: an argument is valid if the truth of its premises guarantees the truth of its conclusion (p. 5). We accordingly analyze the truth conditions of premises and conclusions to determine what is valid. On this basis, our experiments show that nearly everyone is likely to make logical errors. Our analyses of truth conditions may be flawed, and indeed Cohen challenges our account of conditionals (see below). Such challenges, however, are normal in scientific criticism. We grant that if the semantic principle of validity is wrong, then our argument collapses – it is doubtful whether any argument about any topic could be usefully pursued!

**Luchins & Luchins** argue that untrained individuals draw redundant and contradictory conclusions, and that we overestimate their ability. People do make mistakes, but competence must not be confused with performance. Following Chomsky (1965), we treat a specification of competence as an idealization, and we assign the task of explaining errors to the algorithmic theory: "The algorithmic theory . . . should explain the characteristics of human performance – where it breaks down and leads to error, where it runs smoothly, and how it is integrated with other mental abilities" (p. 17). Our subjects make many deductive errors, but they rarely violate the standards of parsimony and novelty. Their valid conclusions never throw semantic information away by introducing new disjunctive alternatives.

**Savion** characterizes the model theory as too idealistic, and questions whether the semantic principle of validity is part of human competence, because she has found it difficult to teach to college students. This claim is akin to arguing against Marr's (1982) theory of vision because it is difficult to teach to students. Marr postulates mental models as the end product of vision, and he too might be accused (at least by Savion) of presupposing an enormous core of competence. Yet, in both his case and ours, it is possible to construct computer programs implementing the theory. The moral is that Savion overlooks the distinction between making a valid deduction and knowing what one is doing (p. 19, p. 147). Rats can make transitive inferences, according to **Davis,** but like Savion's students they probably do not know what they are doing.

**R2. Images, mental models and logical models.** According to the algorithmic theory, reasoners construct models of premises and search for alternative models that are counterexamples to their conclusions. We now consider four aspects of the theory: the relations between mental models and images, the relations between mental models and their counterparts in logic (the model structures of Tarski), the search for counterexamples, and the claim that the theory is a set of separate microtheories.

The mental-model theory has its origin in the introspections of some subjects carrying out syllogistic inferences: their reports of using "images" fitted Craik's (1943) theory of thinking. Subjects sometimes report that they reason "verbally," but they never report using formal rules of inference. We do not reject introspective evidence (*pace* **Braine**), but likewise we do not regard it as sufficient to eliminate formal rule theories. We were remiss in not pointing out the theory's kinship to the ideas of Max Wertheimer, one of the founders of Gestalt psy-

chology, and we thank **Luchins & Luchins** for pointing out the resemblances.

Images are a special case of models (p. 39; see also Johnson-Laird 1983, Ch. 7; *pace* **Stenning & Oberlander** and **ter Meulen**). We hold this position for three reasons. First, there are no detectable differences in performance between those who claim to use images and those who do not. Second, manipulations of imageability have no reliable effects on deduction (p. 140; Johnson-Laird et al. 1989; Newstead et al. 1982; Richardson 1987), although **MacLennan** points out that it is difficult to design decisive studies. Third, our evidence suggests that models contain tokens representing negation (p. 130). Negation cannot be visualized: Signs representing negation can be visualized, but the essential work is done not by the sign but the procedure for interpreting it. Similar annotations can be invoked in responding to **Over,** who argues for the need to distinguish between assumptions and beliefs (of varying degrees of strength). Models represent propositional content and contain separate annotations representing epistemic attitude. **MacLennan** suggests that the annotation representing negation should be treated as an intention ("denial that") toward an *image* (see also **Inder**), and that a disjunction of images works as well as a disjunction of models. The snag is that models can represent many assertions that cannot be readily visualized, for example, "all men are mortal." **MacLennan,** however, argues that the distinction between images and abstract models is one of degree rather than kind. We accept his point that presence in consciousness is a matter of degree for both models and images (see Yates 1985). The content of a model may be available to consciousness, but the process of inference and the format of mental representations are never fully accessible (*pace* **Braine** and **Wetherick**). If they were, then introspection alone would resolve most controversies about mental representation.

What is the relation between mental models and models in the logical sense that Tarski made famous? The question is raised by **Barwise, Bundy, Inder,** and **ter Meulen.** Barwise and Bundy emphasize that human beings cannot construct models in the Tarskian sense. We agree: there are too many possible models (p. 16, p. 36). And Bundy's cautionary tale about his reasoning program is also valid: programs must exploit rule-like maneuvers. Yet, he too respects the distinction in logic between proof theory and model theory. He takes the mental-model theory to imply that a corresponding computer program would give meanings to computational states. He proposes cogent arguments against this consequence; we accept them. The application of the adjective "semantic" to any existing computer program *is* an oxymoron. Models in programs, however, should not be confused with models in minds. Mental models can genuinely represent the world and the meaning of discourse because of their causal relations to the world (p. 213; Johnson-Laird 1983, p. 399 et seq.). Hence, as Inder says, they can reasonably be claimed to be semantic.

We wrote that a mental model functions like a representative sample from the set of possible Tarskian models of a statement (p. 36). This claim is wrong once negation is introduced into models: as **Inder** points out, one model then represents many states. **Barwise** proposes a subtle but better way to construe the psychological theory: a mental model represents a *class* of Tarskian models. This proposal makes no difference to the empirical content of the theory, but it clarifies its formulation. Consider an example akin to Barwise's:

> The triangle is on the right of the star.
> The star is on the right of the line.
> The star is on the right of the circle.

It follows that:

> The triangle is on the right of the circle.

but it does not follow that:

> The line is on the right of the circle.

As Barwise says, reasoners searching for a Tarskian model that refutes the second conclusion may succeed, and so know that the inference is invalid, but those searching for a Tarskian model that refutes the first conclusion will, of course, never succeed, and so must ultimately abandon the search. They will be right to do so, but they cannot know that they are right: they cannot know that the inference is valid. According to the model theory, however, the premises yield the model:

This represents an infinite class of possible Tarskian models in which distances, sizes, shapes, and so on, can all vary. There are only finitely many rearrangements of the objects in a model of this sort, and so they can be examined exhaustively (as in our program for spatial inference), and none of them refutes the valid conclusion. The advantage of mental models (representing infinite classes of situations) is that only a finite number need to be explored to validate deductions of this sort.

The search for counterexamples is at the heart of the model theory, but **Polk** remarks that it appears to underlie few predictions. The reason for the apparent anomaly is simple: if the search is properly carried out, it yields a correct response. Subjects often fail to carry it out properly, and so our predictions emphasize the increasing difficulty of deduction as the number of models increases, and the likelihood of errors based on a proper subset of the possible models. Hence, the predictions *are* founded on the search for counterexamples.

**Bara** points out that even adults often fail to search for counterexamples, and that it took years for Popper's "falsification" criterion for demarcating scientific hypotheses to prevail over verificationism. The need to search for counterexamples is not obvious. In fact, it is not a self-conscious principle for the typical thinker. The best evidence for its existence is that individuals respond "no valid conclusion" reliably better than chance to premises that do not validly yield informative conclusions (see also Oakhill & Johnson-Laird 1985a). And **Barwise** makes a point that we did not exploit: people can know that a conclusion does not follow validly, but this knowledge cannot be accounted for by formal rules.

**Fetzer** believes that there is a crucial equivocation in our theory: people might think there are no counterexamples when in fact there are, and so they might believe that arguments are valid when they are invalid, or that they are invalid when they are valid. Exactly! Individuals make both sorts of error: they fail to find a counterexample, and they fail to see that a set of models supports a common conclusion. But there is no equivocation: the theory

accounts for both sorts of error and for valid arguments, namely, when no counterexamples exist in a domain with a finite number of mental models.

To accommodate a new logical term it is only necessary to describe its meaning, that is, its contribution to the construction of models. The standard inference procedure can then take over (p. 127). Hence, the development of the theory has been piecemeal. **Polk,** while commending the completeness and accuracy of the theory, suggests that it is really a set of microtheories based on a common framework. He lists a set of assumptions, which he claims are not part of the core theory but of particular microtheories. In fact, all of the assumptions that Polk lists *are* part of the core theory:

1. The greater the number of its atomic propositions, the harder an inference: as the number increases so does the size of the models.

2. Individuals reason only about those items that are explicit in their models. This assumption applies to any form of reasoning (see **Legrenzi & Sonino**).

3. Reasoners consider multiple models of premises even if they all support the same conclusion (more on this point below).

4. The need to integrate premises by bringing their referents in common into contiguity applies to any form of reasoning.

5. "Negative" deductions, which call for the deduction of an inconsistency between one model and another, are always harder than affirmative deductions.

6. It is easier to reason from the hypothesis that an assertion is true than the hypothesis that it is false, because it takes work to construct the complement of the set of models representing a hypothesis.

Nevertheless, **Polk** is right about the origins of the theory. We were guided by the general framework in formulating a semantics for each logical domain, and these accounts *are* microtheories, that is, we can formulate them in many ways, which are independent of the inferential mechanism (the search for counterexamples). Otherwise, the components of the theory that appear to be local to particular domains turn out to be general.

**R3. Propositional reasoning.** An exclusive disjunction, such as "Either there is a circle or else there is a triangle, but not both," has the following initial models:

[○]
[△]

where the brackets indicate that a proposition has been exhaustively represented. The procedure for fleshing out models works as follows: when a proposition has been exhaustively represented in one or more models, it adds its negation to any other models. The procedure adds that there is *not* a triangle to the first of the two models above because triangles are exhausted in the second model. The final result of fleshing out the models is:

[○] [¬△]
[¬○] [△]

where "¬" denotes negation. These two models correspond to those rows in the disjunction's truth table that are true. In general, the number of *explicit* models for a propositional deduction equals the number of rows that are true in a truth table of all the premises. Exhaustion is therefore a device that allows the inferential system to represent certain information implicitly – it can be made explicit but at the cost of fleshing out the models. **Andrews** grasps the basic idea but wonders how it applies to complex propositions; **Hodges** gives up on it and, it seems, on the model theory as a whole; perhaps the problems of psychology are too murky for him after the clarity of logic, though we agree with him that cognitive scientists from different disciplines should talk to one another. Andrews asks how exhaustion applies to the representation of the sentence:

Either there is a circle or a triangle, or a triangle and a square, but not both.

This is a reasonable question from a logical point of view, but it misses the psychological point. The answer is that exhaustion could be used recursively. The procedure represents the main connective (A or B, but not both) first:

[A]
[B]

It then represents proposition A (there is a circle or a triangle) which we will assume to be an exclusive disjunction:

[○]
[△]

and proposition B (there is a triangle and a square):

[△ □]

Finally, it substitutes these models in place of A and B in the initial model:

[[○]
[△]
[△ □]]

Fleshing out can be accomplished in a similarly recursive way. The overall models become:

A ¬B
¬A B

The model corresponding to B is already fleshed out, and the model corresponding to A fleshes out as:

○ ¬△
¬○ △

A and B (and their complements) yield the final explicit models:

A ¬B yields: ○ ¬△ □
○ ¬△ ¬□
¬○ △ ¬□
and ¬A B yields: ○ △ □

Even here, recursive exhaustion is too complicated to be psychologically plausible. Our study of "double disjunctions" shows that individuals have difficulty in representing three distinct models (p. 56).

**R4. Conditionals.** Our theory of indicative conditionals combines Grice's (1975) conversational implicatures with implicit representations. The antecedent and consequent are represented explicitly in one model, and there is an alternative wholly implicit model. These initial models yield judgments corresponding to a "defective" truth

table, difficulty with *modus tollens*, and difficulty with Wason's (1966) selection task. If the models are fleshed out, they yield a truth-functional interpretation. Two reviewers express qualms about this account. **Over** argues that an indicative conditional tends to be asserted and accepted when its consequent seems highly probable given its antecedent. We give no account of this reading, because we doubt it. At the time of writing, for example, we believe that, given that Bush is nominated, Clinton is highly likely to win, but we would *not* accept, assert, or consider to be true, the following conditional:

If Bush is nominated, Clinton will win

though we would accept:

If Bush is nominated, then Clinton is highly likely to win.

Hence, our doubts over Over's claim.

**Cohen** cites the following strange inference (see also Cooper 1968), which he regards as problematic for our analysis:

If John's automobile is a Mini, John is poor, and, if John's automobile is a Rolls, John is rich.
Therefore, Either, if John's automobile is a Mini, John is rich (sic), or, if John's automobile is a Rolls, John is poor (sic).

The inference is valid, though it seems not to be. However, it throws semantic information away with a vengeance. Even assuming that John cannot be both rich and poor, the conclusion has eleven models. We claim that hardly anyone can mentally envisage these explicit models for a *disjunction* of the two constituent conditionals. Taken individually, the conditionals conflict with those in the premises, and so the inference seems invalid.

We reported a study in which subjects balked at *modus ponens* (p. 81 et seq; Byrne 1989). When we gave them such premises as:

If Paul goes fishing, then Paul has a fish supper
If Paul catches some fish, then Paul has a fish supper
Paul goes fishing

they do not draw the conclusion:

Paul has a fish supper

contrary to the claim that formal rules are *automatically* applied to any assertions of the appropriate logical form. Several reviewers objected to our referring to the "suppression" of *modus ponens*, and suggested instead that subjects reject one of the premises (**Bach**), or take the second conditional to render the first one uncertain (**Over**) or false (**Savion**; see also Politzer & Braine 1991; and for a response Byrne 1991). Our claim, like **Grandy**'s, is that subjects interpret the two conditionals as equivalent to "if Paul goes fishing and catches some fish, then Paul has a fish supper." They produce this sort of paraphrase of the two conditionals and they do not say that one conditional renders the other false (Byrne & Johnson-Laird 1992). In the light of the reviewers' claims, Byrne and Handley carried out an experiment in which the subjects explicitly judged the truth values of the conditionals: they tended to judge that both conditionals were true, especially after they had carried out the inferential task! **Fillenbaum**'s judicious discussion leads him to a conclusion with which we concur: the interpretative component is critical. Rule theories need to explain the recovery of logical form, and the model theory needs to explain how background knowledge produces models corresponding to those of the paraphrase.

The model theory yields an explanation of performance in Wason's selection task (pp. 75–81). It is the only account that purports to explain the effects of all the variables, including abstract and realistic materials. **Evans** regards it as perfunctory, but does not explain why. **Green** correctly points out that the critical issue concerns what factors cue subjects to flesh out their initial models. He reports a study that confirmed the core of the model theory, but some subjects who envisaged the critical counterexample failed to select it. There are possible reasons for such a failure, such as the lability of the identification. We take comfort in the overall correlation between identifying counterexamples and making correct selections.

**Manktelow** points out a real difficulty for our account, but in fact it is a problem for any theory. Insightful subjects should select all four cards if they have made an equivalence interpretation of the conditional. Perhaps they *would* select all four cards if the experimental procedure did not imply that this selection was somehow redundant. Manktelow also argues that we need to distinguish between evaluations of truth value and evaluations of violations of rules, and that we need a procedure to convert (invalidly) the conditional to a "q only if p" form in order to explain ¬p and q selections. In fact, we do not propose a verbal conversion of "if p then q" to "q only if p": we argue that the interpretation of the rule in context yields the "only if" models. However, we entirely accept that there is a difference between judging truth values and judging violations. The theory's formulation in the latter case should read: select those cards that have a bearing on compliance or violation of the rule. We also accept that preferences and "point of view" matter (cf. Johnson-Laird & Byrne 1992). Our only doubt is whether preferences are based on immediately available "utilities." What, for example, is the utility to us of Manktelow's commentary? Certainly we would give it a positive value, but to obtain a utility in the classical decision-theoretic sense would be difficult, and would itself require us to reason.

**Pollard** remarks that the rules yielding the most insight into the selection task are those for which counterexamples are known to the subjects. Hence, he urges us to adopt a "nonlogical" model of the conditional that represents explicit counterexamples. We accept this recommendation, but with one qualification: The counterexample is not part of the models for the conditional but is represented as impossible (given a true rule) or as impermissible (given a deontic rule). Indeed, we hinted at this very idea (p. 80). To represent the conditional as embracing the counterexample would be a little too "nonlogical"!

**R5. Syllogisms.** Commentators raised more questions about our account of syllogistic inference than about any other topic. The most serious question was: Is there an objective procedure for generating the models? We have implemented the theory in a computer program; it differs in detail from the theory in Chapter 6, but it yields the same general predictions. The principle for combining models is, as **Garnham** surmises, to combine those indi-

viduals sharing the property denoted by the middle term (having made sure that there are the same numbers of them). This assumption leads to some initial models that support invalid conclusions, which are crucial for the explanatory power of the theory. Even without a program, however, it is important to realize that there is a simple objective test of the model theory for any domain of deduction: erroneous conclusions should be consistent with the truth of the premises, that is, the subjects are considering some, but not all, of the possible models of the premises and so they draw conclusions that are possibly true rather than necessarily true. This prediction cannot be made by the formal rule theories.

As **Garnham** points out, the following sort of premises:

All the A are B
All the B are C

are consistent with eight distinct sorts of situation (assuming existential import). They all contain individuals with the three properties:

a b c

The presence or absence of each of three other sorts of individual produces the eight possibilities:

¬a b c

and:

¬a ¬b c

and:

¬a ¬b ¬c

Four of the possibilities correspond to Euler circles for the premises: Euler circles do not represent the presence or absence of individuals with none of the three properties. How is it possible for untrained reasoners to do so well with these premises? The answer is that reasoners' models capture the possibilities implicitly. In an earlier version of the theory, models could represent individuals that might, or might not, exist (Johnson-Laird & Bara 1984). In *Deduction*, we introduced the notion of exhaustiveness for propositional reasoning, and so, granted the relation between conditionals and universal premises, we adopted the same device for universal quantifiers. The premises above accordingly yield the single initial model:

[[a] b] c
[[a] b] c
. . .

The three dots represent implicit individuals with no specific properties, and the brackets indicate exhaustion. The representation of one property is exhausted in relation to another: a's can occur in fleshing out an implicit individual only if that individual is assigned the property denoted by b, and b's can occur in another individual only if that individual is assigned the property denoted by c. Hence, the model can be fleshed out in the eight different ways. Each way corresponds to a class of Tarskian models, but the initial model captures them all.

The initial model also yields the conclusion, "All the C are A," which is an error that subjects sometimes make. As **Newstead** demonstrates, this conclusion is refuted by a second model:

[[a] b] c
[[a] b] c
c
. . .

which supports only the conclusion: "Some of the C are A." In general, where only one model is needed to formulate a valid conclusion, the task should be harder if the conclusion is consistent with a further model that falsifies another conclusion. Hence, as in this example, AA premises yielding valid A conclusions should be harder than IA or AI premises yielding valid I conclusions. (We use the abbreviations: A for assertions of the form "All X are Y," I for "Some X are Y," E for "No X are Y," and O for "Some X are not Y.") The evidence appears to bear out this prediction (Table 6.1). This explanation should help Newstead to explain the effects of belief upon syllogistic inference. It also lays to rest one of **Polk**'s worries. He points out that reasoners are affected by multiple spatial models even if they support the same conclusion, but not similarly affected by multiple syllogistic models. The same difficulties occur in both domains.

**Newstead** and **Pollard** both wonder whether an O premise, "Some of the A are not B," may pragmatically imply the truth of the I premise, "Some of the A are B." One small class of conclusions can be explained in this way. For example, premises of the form:

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

elicited the following percentages of responses (over four experiments with correlated results):

Some of the A are not C (32%)
Some of the C are not A (7%)
No valid conclusion (35%)

These are the responses predicted by the model theory. The remaining responses were almost all:

Some of the A are C (15%)

which may reflect the pragmatic interpretation of the O premise.

**Stenning & Oberlander** have shown how a new method of using Euler circles is equivalent to the model theory. We applaud their ingenuity. They claim that the method is more constrained than mental models, but it is not clear why. Nor is it clear why these authors believe that we rule out graphical methods. Images are a special case of models, and presumably they are graphical. What we do maintain, however, is that traditional Euler circles cannot represent multiply quantified relations, and so they are unlikely to be used by logically untrained individuals, who move freely from singly to multiply quantified relations (pp. 134–35).

**Wetherick** rejects our account of syllogistic reasoning and argues that some subjects are prey to an "atmosphere" effect in which they match their conclusions to the mood of one of the premises, whereas other subjects "engage with the logical task." Some subjects may sometimes draw a conclusion because it matches the mood of a premise, but matching is implausible as a general account of syllogistic reasoning. First, there is a simple alternative explanation: the initial model of any conventional syllogism yields a conclusion matching the mood of at least one premise. Matching, however, cannot explain why sub-

jects ever respond with conclusions that emerge from subsequent models or why they respond, "there is no valid conclusion." Second, a striking failure to demonstrate matching occurred in a study of "only" as a quantifier. When syllogisms contain "only," subjects were most reluctant to draw a matching conclusion, preferring instead the quantifier, "all" (p. 129). Third, matching fails to explain performance with multiply quantified premises (p. 140). Wetherick accuses us of overlooking an alternative explanation for our results. The truth is that we have not overlooked it, but eliminated it. Indeed, matching seems to be a more accurate account of its own origins than of reasoning.

**R6. Multiple quantifiers.** Existential quantifiers should be more difficult than universal quantifiers, **Grandy** argues, just as disjunction is more difficult than conjunction. The hypothesis is aesthetically pleasing; the facts are ugly. The easiest syllogism contains an existential (Some of the A are B, All the B are C), and the easiest doubly quantified problems, as he acknowledges, include those with existentials. Conversely, some difficult deductions do not contain existentials. He points out that proofs can be difficult because of the interactions among the rules for quantifiers. We leave to rule theorists the task of determining whether these interactions could in principle explain our results with multiply quantified deductions. We doubt it. Our untrained subjects generate conclusions in a minute or so; we have yet to see logicians derive new conclusions – not just prove given conclusions – in a comparable time. Either our subjects have a remarkable tacit ability at logical derivations (which is, alas, not available in the logic classroom), or, as we believe, they are reasoning by other means.

**Crawford** argues for an intrinsic logical difficulty of ∀ ∃ quantification, that is, as when the assertion: "All musicians are related to some authors" is interpreted as all musicians are related to some author or other. The reason is that assertions of this form may call for an unbounded set of individuals (see also, pp. 178–80). Once again, the theory is beautiful, but not the facts. We found no differences in reasoning with assertions of the ∀∃ form and the ∃∀ form (pp. 142–43). Crawford suggests that deductions that are computationally tractable but difficult for people may yield evidence about human reasoning algorithms. We agree, but we note that people often work with small-scale problems in intractable domains, for example, they can deduce their own parsimonious conclusions from propositional premises.

**R7. Everyday informal reasoning.** Not only is most reasoning utterly unlike syllogistic or propositional reasoning, claims **Fisher,** but it is not even deductive (or inductive). He and many of his colleagues in the movement for "informal logic" believe that logic has little application to the analysis of everyday arguments (see, e.g., Toulmin 1958). **Galotti & Komatsu** suggest, however, that the model theory may be a fruitful way to explain everyday arguments, and this approach looks promising (see pp. 205–6, and Morton 1988). The model theory shows how to combine valid deduction and the "nonmonotonic" reasoning that occurs in undoing arbitrary or default assumptions (pp. 180–83). And a major implication of the theory is that people are *inferential satisficers* (cf. Simon 1991), that is, once they find a model that fits their beliefs, they do not tend to search for alternatives (p. 126). This negative tendency seems to be the cognitive cause of many disasters, for example, the operators at Three Mile Island thought that the high temperature of a relief valve arose from a leak and overlooked the possibility that the valve was stuck open. The theory makes the same prediction about everyday arguments: people draw conclusions that are true in some models of the premises and often overlook alternative models. Hence, even in daily life, a counterexample is likely to devastate an argument: its effects are explicable only in terms of an underlying deductive competence (*pace* **Fisher,** and **Luchins & Luchins**).

Our theory assumes that general knowledge is used in constructing models, and so it dovetails with Tversky and Kahneman's (1973) "availability" heuristic (p. 206, *pace* **Savion**). Any *available* knowledge can be embodied in a model, and the process of comprehension make some items of knowledge more available than others. This account finesses the mechanism underlying availability (as pointed out by **Green, Inder, Legrenzi & Sonino,** and **Stevenson**). Inder remarks that we are assuming that the brain comes fitted with "mental-model accelerators," and he warns us not to push too much of the explanatory load into cognitive architecture. We take the point, but the retrieval of relevant knowledge is the only major problem for which we have presupposed a solution. Both Green and Stevenson suggest that expertise in a domain may lead to the development of content-specific rules, and **Evans** believes that such rules might even be implemented in a connectionist network (cf. also **MacLennan**). The evidence about implicit reasoning in comprehension suggests that people do construct models (see, e.g., Garnham 1987), and the general knowledge used in their construction could be represented in content-specific rules (Newell 1990; see also multiple book review, *BBS* 15(3) 1992). So far, however, no evidence has identified how knowledge is mentally represented – it could be in the form of rules, assertions, models, or networks.

For nonmonotonic reasoning, **Chater** suggests that the model theory needs to be constrained by a logic, and that reasoning by searching for counterexamples is inappropriate because they always exist. We see no reason why the construction of models needs to be constrained by a logic: it is constrained by available knowledge. We likewise see no difficulty in basing nonmonotonic reasoning on a search for counterexamples. Consider Chater's own example: You infer from the sound of purring that your cat is trapped in the cellar, but you override this conclusion if you catch sight of it in the garden. You withdraw the inference because of a counterexample. And isn't this very example a counterexample to the claim that counterexamples are an inappropriate method of nonmonotonic reasoning? And so shouldn't the claim be withdrawn? Nonmonotonically?

**Oaksford** argues that the model theory is unlikely to yield a tractable account of nonmonotonic reasoning. Perhaps not. But nothing implies that people are using a tractable procedure. Our experiments on everday inference (p. 205) suggest that they bring to mind an available scenario. They think of some alternatives, but they overlook many possibilities. They may be using an intractable algorithm that is defeated by the magnitude of the task.

The model theory accounts for the undoing of arbitrary or default assumptions, but it does not explain another sort of nonmonotonic reasoning. When you draw a conclusion that clashes with your beliefs, you have to reconcile the discrepancy. You may revise your inference, your premises, or your beliefs. You search, as **Bach** emphasizes, for the best explanation (Harman 1986). How you do so is largely unknown, but you may create new ideas. If we said that we had solved this problem (and **Chater** takes us to have said so), then we were mistaken. It remains deeply mysterious.

**R8. Rationality versus relativism.** We rejected relativism and defended universal rationality founded, not on formal rules of inference, but on the semantic principle of validity. **Engel** largely agrees with us, but argues for two sorts of rationality: rationality of purpose, and rationality of process (see Evans, in press). We accept the distinction, though we worry about its psychological basis – how, for example, are the two sorts of rationality mentally embodied, and how do they interact? Engel suggests that mental models may play a part in rationality of purpose, and so perhaps they provide an underlying common framework. He also defends the process of "reflective equilibrium" that enables people to bring their intuitions and normative principles into harmony. People may go through such a process, but our concern is the basis of their intuitions, not their normative principles.

**Braine** considers the occurrence of logical terms in all human languages as more naturally explained by a mental logic than by the principle of semantic validity. It is difficult to advance beyond the mere trading of intuitions. However, relativists argue that no account of deductive competence justifies a unique system of logical rules (pp. 208–9), and Braine proposes no solution to this problem.

Only one commentator hints at a defense of relativism. **MacLennan** suggests that we should allow for culturally permissible inconsistencies. We doubt whether anyone would knowingly accept an inconsistency on the grounds that it is culturally acceptable – except perhaps as a Whitmanesque posture: "Do I contradict myself? Very well then I contradict myself. (I am large, I contain multitudes.)" Certain inconsistencies may be invisible to a culture, but do they remain so once some intrepid individual points them out? What matters is the recognition of inconsistencies, because consistency *is* a universal of logic (*pace* MacLennan). And so, we believe, is the semantic principle of validity.

**R9. Extensions of the mental-model theory.** Numerous commentators who accept the core of our argument raise the possibility of extending the model theory to cope with other mental processes, including, as **Bach** and **Galotti & Komatsu** suggest, those for which formal rules would be implausible. **Baron** argues that a modest generalization of the theory can describe all goal-directed thinking: thinking is a search for possibilities, evidence, and goals. Search is biassed towards positive evidence, he says, particularly because some people do not understand the importance of negative evidence. Another factor arises directly from models (p. 79): as **Legrenzi & Sonino** point out, individuals focus on what is explicit in their models and neglect what is only implicit in them. **Tweney** remarks that most scientific thought fits comfortably within the model framework, but he rightly points to the problem of how such models are initially created by scientists (see Johnson-Laird, 1992, for some thoughts on this issue). What is striking is the role that the manipulation of models appears to play in scientific innovations (see, e.g., Wise 1979). **Bara** argues that the model theory needs to be extended to the development of deductive ability in children. The beginnings of such an extension can be found in Johnson-Laird (1990b): intellectual growth is not the development of new mental operations (*pace* Piagetians) but the development of new concepts – a hypothesis urged by Carey (1991), Keil (1991), and others. Verbal instruction is no substitute for the construction of models of the world.

**Davis** draws attention to the burgeoning literature on animal reasoning and asks whether the model theory has a role to play here. His results are presumptive evidence that rats form an internal representation that enables them to make transitive inferences. [See also Davis & Pérusse: "Numerical Competence in Animals" *BBS* 11(4) 1988.] Such a representation could take many forms, but, as he says, it is unlikely to depend on a postulate of transitivity and the predicate calculus. Animals build up spatial models of their environment and combine models that have been separately learned. Because models probably owe their origin to the evolution of perceptual ability (Marr 1982), they are likely to play a role in the inferences of humans and nonhumans alike. What is unique to humans is natural language and, perhaps, the ability to reflect on their own performance.

**R10. The incompleteness of the research.** We have almost completed our reply to the commentators' detailed queries, but we will consider some methodological criticisms before we review the current status of the theory. **Ter Meulen** has reservations about our experiments: she accuses us of not manipulating the order of premises, and of using materials of a limited genre. **Evans** also claims that our studies, except for those on belief-bias, use materials with an arbitrary content. These claims are incorrect, though we could not describe all our experimental manipulations in a research monograph – a limitation for which readers should be profoundly grateful. We did, however, report one major effect of order of premises (Table 6.2; for other effects, see e.g., Erhlich & Johnson-Laird 1982; Legrenzi et al. 1992), and we used realistic materials in our studies of the "suppression" of *modus ponens* (pp. 81–82; Byrne 1989), paraphrasing with conditionals (pp. 84–85; Byrne & Johnson-Laird 1992), and informal everyday inferences (pp. 205–6). The major studies of formal rule theories have used arbitrary materials, and for purposes of comparison we used them too.

As we wrote in *Deduction*, the model theory is incomplete (p. 213). It does not explain, for example, how the search for counterexamples is carried out; our experiments on the topic were not successful. Psychologists are familiar with this problem, and, where there are no grounds for theorizing, do not demand that a theory be specified to the last detail. Our colleagues in other disciplines, however, chide us about the theory's incompleteness:

What counts as a mental model?

What other forms of mental representation are there apart from models?

How does the theory apply to nonstandard quantifiers or to modal reasoning?

What drives the search for alternative models?

How do people know when no more models are needed?

Do two people have the same mental models if they make the same inferences?

We have tried to answer the first three questions (Johnson-Laird 1983), but we do not know the answers to the rest.

Some reviewers found our account incomplete in at least one way in which it is not (**Bach, Falmagne, ter Meulen**). They ask: How are models constructed from premises? We described a compositional semantics for connectives, relations, and quantifiers; and we showed how it is implemented in computer programs (see Ch. 9). We can provide more detail if anyone requires it. **Fetzer** remarks that to account for all aspects of performance places too much of a demand on the theory: performance can be affected by motives, beliefs, ethics, ability, and other factors. A complete theory of the mind would aim to embrace such effects, and our theory aims to account for effects of beliefs and ability.

The question of individual differences drew comment from **Bach** and **ter Meulen.** Bach asks whether some "hard" problems might be easy for some people. The answer is that the differences between moderately hard problems and easy ones can disappear for exceptional reasoners, but we have never observed reliable reversals in difficulty. The trends are remarkably robust, for example, we have never tested anyone who does not do better with one-model syllogisms than with multiple-model syllogisms. Model-based accounts of individual differences are now under way, especially as a result of Polk's studies and his development of software for fitting parameterized theories to individuals' data (Polk & Newell 1992). Why, Bach asks, are some people better at deduction than others? The answer, judging by the patterns of systematic error, is that there do not appear to be vast differences either in models or strategies. We know that a measure of the capacity of working memory accounts for part of the variance, and that a measure of the ability to perceive what is common to different drawings accounts for rather more (Bara et al. 1992). But we do not know how best to interpret these correlations, and we are far from a causal account of differences in ability.

Finally, some reviewers believe our theory is incomplete because it does not include formal rules. They urge us to combine models and rules. They particularly want the rule for *modus ponens:* **Pollard** asserts that without it individuals would have difficulty constructing models for conditionals, **Falmagne** argues that it is required for rapid automatic inferences relying on form, and **Wetherick** claims that no theory applying general principles to particular cases can avoid it. In fact, models for conditionals can be constructed without the rule for *modus ponens:* our computer programs construct them solely from the "truth conditions" of the connective. People make *modus ponens* deductions rapidly, but it does not follow that they rely on a formal rule to do so – even if the content of the premises is abstract or based on nonce words. They could build models containing abstract or nonce items (*pace* **Savion**). And unified theorem-provers apply general principles to particular cases without using *modus ponens*: they rely instead on unification and the resolution rule (see pp. 26–27 for a brief account). With experience in a reasoning task, as **Stevenson** suggests, subjects may begin to construct formal rules for themselves (see p. 202; Galotti et al. 1986). But few individuals seems capable of formulating rules that capture all the valid deductions they can make (as Victoria Shaw has found in an unpublished study).

The model theory proposes that variables occur in the initial semantic representations of sentences (pp. 171–73), and that they are then instantiated in models as small finite sets of tokens corresponding to individuals (pp. 177–80). **Braine** correctly claims that models do not contain variables, but he overlooks their occurrence in the initial semantic representations (pp. 171–73). **Ter Meulen** understands that our programs use variables in this way, but does not seem to realize that the programs are implementations of our theory.

**R11. Conclusions.** No commentator proposes a new theory accounting for the phenomena of deductive reasoning. So: mental models or formal rules? Or both? A theory that combines both is probably irrefutable, that is, no evidence could ever show it to be false. We therefore preferred to exercise parsimony and to reject the psychological theories based on formal rules. **Braine** claims that the argument for mental models rests solely on parsimony; and **Fetzer** claims that it rests solely on the effects of content. They overlook the real case. It is the experimental evidence as a whole. There are psychological theories based on formal rules for relational reasoning, but the empirical evidence counts against them. There are such theories for propositional reasoning, but they fail to explain differences in difficulty, systematic errors, and the effects of content. No such theories exist for syllogisms or multiply quantified deductions. In each of these domains, the model theory explains the phenomena.

Where do we go from here? Our immediate tasks are to develop better accounts of everyday reasoning, to explain how reasoners discover new strategies in metareasoning (see Byrne & Handley 1992), to interrelate reasoning and decision making, and to find out how diagrams improve reasoning – especially in the light of recent studies (Barwise & Etchemendy 1991; Bauer & Johnson-Laird 1993). We thank the commentators for helping us to strengthen the exposition of the theory, to clarify the relations between mental models and Tarskian models, and to pursue extensions of the theory into other realms of thought.

## References

**Letters *a* and *r* appearing before authors' initial refer to target article and response respectively.**

Adams, E. W. (1975) *The logic of conditionals.* Reidel. [DO]

Anderson, C. A. (1982) Inoculation and counterexplanation: Debiasing techniques in the perseverance of social theories. *Social Cognition* 1:126–39. [JBaro]

Anderson, J. R. (1982) Acquisition of cognitive skill. *Psychological Review* 89:369–406. [RJS]

Arkes, H. R., Faust, D., Guilmette, T. J. & Hart, K. (1988) Eliminating the hindsight bias. *Journal of Applied Psychology* 73:305–7. [JBaro]

Bach, K. (1970) Part of what a picture is. *British Journal of Aesthetics* 10:119–37. [KB]

(1984) Default reasoning: Jumping to conclusions and knowing when to think twice. *Pacific Philosophical Quarterly* 65:37–58. [KB]

Bara, B. G., Bucciarelli, M. & Johnson-Laird, P. N. (1992) The development of syllogistic reasoning. Unpublished paper. Center for Cognitive Science, University of Turin. [rPNJ-L]

Barnes, B. & Bloor, D. (1982) Relativism, rationalism, and the sociology of knowledge. In: *Rationality and relativism*, ed. M. Hollis & S. Lukes. Basil Blackwell. [aPNJ-L, PE]

Baron, J. (1988) *Thinking and deciding*. Cambridge University Press. [JBaro]
(1991) Beliefs about thinking. In: *Informal reasoning and education*, ed. J. F. Voss, D. N. Perkins & J. W. Segal. Erlbaum. [JBaro]

Baron, J., Badgio, P. & Gaskins, I. W. (1986) Cognitive style and its improvement: A normative approach. In: *Advances in the psychology of human intelligence*, vol. 3, ed. R. J. Sternberg. Erlbaum. [JBaro]

Baron, J., Badgio, P. & Ritov, Y. (1991) Departures from optimal stopping in an anagram task. *Journal of Mathematical Psychology* 35:41–63. [JBaro]

Barwise, J. (1986) Conditionals and conditional information. In: *On conditionals*, ed. C. Ferguson, J. Reilly, A. ter Meulen, & E. C. Tragott. Cambridge University Press. [AGBtM]
(1989) *The situation in logic*. CSLI Lecture Notes, Stauford University. [AGBtM]

Barwise, J. & Etchemendy, J. (1991) Visual information and valid reasoning. In: *Visualization in teaching and learning mathematics*, ed. W. Zimmermann & S. Cunningham. M.A.A. Notes #19, Mathematical Association of America. [rPNJ-L, J Barw, KS]

Barwise, J., Gawron, J. M. & Tutiya, S., eds. (1990) *Situation theory and its applications*, vol. 2. CSLI Lecture Notes, Stanford University. [AGBtM]

Barwise, J. & Perry, J. (1983) *Situations and attitudes*. Bradford Books/MIT Press. [AGBtM]

Baner, M. I. & Johnson-Laird, P. N. (1993) How diagrams can improve reasoning. *Psychological Science* (in press). [rPNJ-L]

Beattie, J. & Baron, J. (1988) Confirmation and matching bias in hypothesis testing. *Quarterly Journal of Experimental Psychology* 40A:269–97. [JBaro]

Begg, I. & Denny, J. (1969) Empirical reconciliation of atmosphere and conversion interpretations of syllogistic reasoning. *Journal of Experimental Psychology* 81:351–54. [aPNJ-L]

Bereiter, C. & Scardamalia, M. (1989) Intentional learning as a goal of instruction. In: *Knowing, learning and instruction: Essays in honour of Robert Glaser*. Erlbaum. [RJS]

Berkeley, G. (1710/1965) *A treatise concerning the principles of human knowledge*. Bobbs-Merrill. [KB]

Berry, D. C. & Broadbent, D. E. (1984) On the relationship between task performance and associated verbalizable knowledge. *Quarterly Journal of Experimental Psychology* 36A:209–31. [JStBTE]

Beth, E. W. (1955/1969) Semantic entailment and formal derivability. *Mededelingen van de Koninklijke Nederlandse Akademie van Wetenschappen, Afdeling, Letterkunde*, N. R. 18:309–42. Reprinted in Hintikka, J. (ed.) *The philosophy of mathematics*. Oxford University Press, 1969. [aPNJ-L, ADA, JHF]

Beth, E. W. & Piaget, J. (1966) *Mathematical epistemology and psychology*. Reidel. [aPNJ-L]

Blumberg, A. (1967) Logic, modern. In: *The encyclopedia of philosophy*, ed. P. Edwards. Macmillan and The Free Press. [JHF]

Braine, M. D. S. (1978) On the relation between the natural logic of reasoning and standard logic. *Psychological Review* 85:1–21. [aPNJ-L]
(1990) The "natural logic" approach to reasoning. In: *Reasoning, necessity, and logic: Developmental perspectives*, ed. W. F. Overton. Erlbaum. [MDSB]
(1992) Approaches empiriques dn langage de la pensée. In: *Epistomologie et cognition*, ed. D. Andler, P. Jacob, J. Proust, F. Récanati & D. Sperber. Pierre Mardaga. [MDSB]
(in press) The mental logic and how to discover it. In: *The logical foundations of cognition*, ed. J. Macnamara & G. E. Reyes. Oxford University Press (Oxford). [MDSB]

Braine, M. D. S., Reiser, B. J. & Rumain, B. (1984) Some empirical justification for a theory of natural propositional logic. In: *The psychology of learning and motivation: Advances in research and theory*, vol. 18, ed. G. H. Bower. Academic Press. [aPNJ-L, MDSB, KMG]

Broadbent, D. E., Fitzgerald, P. & Broadbent, M. H. P. (1986) Implicit and explicit knowledge in the control of complex systems. *British Journal of Psychology* 77:33–50. [JStBTE]

Brown, A. L., Campione, J. C. & Day, J. D. (1981) Learning to learn: On training students to learn from texts. *Educational Researcher* 10:14–21. [RJS]

Brown, A. L. & Day, J. D. (1983) Macrorules for summarising texts: The development of expertise. *Journal of Verbal Learning and Verbal Behaviour* 22:1–14. [RJS]

Brnner, J., Greenfield, P. & Oliver, R. (1966) *Studies in cognitive growth*. Wiley. [BGB]

Bryant, P. E. & Trabasso, T. (1971) Transitive inferences and memory in young children. *Nature* 232:456–58. [HD]

Bundy, A. (1973) Doing arithmetic with diagrams. In: Proceedings of the Third International Joint Conference on Artificial Intelligence, ed. N. Nilsson. (Also available from Edinburgh as DCL Memo No. 61) [AB]

Burke, P. (1986) Strengths and weaknesses in the history of mentalities. *History of European Ideas* 7. [aPNJ-L]

Byrne, R. M. J. (1989) Suppressing valid infererences with conditionals. *Cognition* 31:61–83. [arPNJ-L, SF, DO]
(1991) Can valid inferences be suppressed? *Cognition* 39:71–78. [rPNJ-L, SF, DO]

Byrne, R. M. J. & Handley, S. J. (1992) Meta-deductive strategies. Uupublished paper. Department of Psychology, Trinity College, University of Dublin. [rPNJ-L]

Byrne, R. M. J. & Johnson-Laird, P. N. (1989) Spatial reasoning. *Journal of Memory and Language* 28:564–75. [aPNJ-L]
(1990) Remembering conclusions we have inferred: What biases reveal. In: *Cognitive biases: Their contribution for understanding human cognitive processes*, ed. J.-P. Caverni, J.-M. Fabre & M. Gonzalez. North-Holland. [aPNJ-L]
(1992) The spontaneous use of propositional connectives. *Quarterly Journal of Experimental Psychology* 45A:89–110. [rPNJ-L]

Carey, S. (1991) Knowledge acquisition: Enrichment or conceptual change? In: *The epigenesis of mind: Essays on biology and cognition*, ed. S. Carey & R. Gelman. Erlbaum. [rPNJ-L]

Carroll, L. (1895) What Achilles said to the tortoise. *Mind* 4:278–80. [KB]

Ceci, S. J. & Liker, J. (1986) Academic and nonacademic intelligence: An experimental separation. In: *Practical intelligence: Nature and origins of competence in the everyday world*, ed. R. J. Sternberg & R. K. Wagner. Cambridge University Press. [RJS]

Chater, N. & Oaksford, M. (1993) Logic, mental models and everyday reasoning. *Mind and Language* (in press). [NC/MO]

Cheng, P. W. & Holyoak, K. J. (1985) Pragmatic reasoning schemas. *Cognitive Psychology* 17:391–416. [aPNJ-L, DWG, KIM]

Cheng, P. W., Holyoak, K. J., Nisbett, R. E. & Oliver, L. M. (1986) Pragmatic versus syntactic approaches to training deductive reasoning. *Cognitive Psychology* 18:293–328. [KMG]

Chi, M. T. H. (1978) Knowledge structures and memory development. In: *Children's thinking: What develops?* ed. R. Siegler. Erlbaum. [RJS]

Chomsky, N. (1965) *Aspects of the theory of syntax*. MIT Press. [rPNJ-L, BGB]

Churchland, P. S. (1986) *Neurophilosophy*. MIT Press. [aPNJ-L]

Clark, H. H. (1969) Linguistic processes in deductive reasoning. *Psychological Review* 76:387–404. [aPNJ-L]

Clifton, C., Jr. & Ferreira, F. (1987) Modularity in sentence comprehension. In: *Modularity in knowledge representation and natural-language understanding*, ed. J. Garfield. MIT Press. [KMG]

Cohen, L. J. (1981) Can human irrationality be experimentally demonstrated? *Behavioral and Brain Sciences* 4(3):317–70. [LJC, PE]

Collins, A. & Michalski, R. (1989) The logic of plausible reasoning: A core theory. *Cognitive Science* 13:1–49. [aPNJ-L]

Conan Doyle, A. (1892) *A study in scarlet*. Harper. [aPNJ-L]

Cooper, R., Mukai, K. & Perry J., eds. (1990) *Situation theory and its applications*, vol. 1. CSLI Lecture Notes, Stanford University. [AGBtM]

Cooper, W. S. (1968) The propositional logic of ordinary discourse. *Inquiry* 11:295–320. [rPNJ-L]

Cosmides, L. (1989) The logic of social exchange: Has natural selection shaped how humans reason? Studies with the Wason selection task. *Cognition* 31:187–276. [aPNJ-L, DWG, KIM, PP]

Craik, K. (1943). *The nature of explanation*. Cambridge University Press (Cambridge). [arPNJ-L, PE]

Crawford, J. M. & Kuipers, B. J. (1991) Negation and proof by contradiction in access-limited logic. *Proceedings of the Ninth National Conference on Artificial Intelligence*. AAAI Press/MIT Press. [JMC]

Cummins, R. (1986) Inexplicit information. In: *The representation of knowledge and belief*, ed. M. Brand & R. M. Harnish. University of Arizona Press. [KB]

Davis, H. (1992) Transitive interference in rats. *Journal of Comparative Psychology* (in press). [HD]
(in press) Attributing consciousness to animals: Is it logical or necessary? In: *Anthropomorphism, anecdotes and animals*, ed. R. W. Mitchell, N. S. Thompson & H. L. Miles. University of Nebraska Press. [HD]

Donald, M. (1991) *The origins of the modern mind: Three stages in the evolution of culture and cognition*. Harvard University Press. [PE]

Duncker, K. (1926) A qualitative (experimental and theoretical) study of

productive thinking (solution of complicated problems). *Journal of Genetic Psychology* 33:642–708. [ASL]
(1945) On problem solving. (Translated by L. S. Lees). *Psychological Monographs* 58, no. 5. (Originally published in 1935.) [ASL]
Ehrlich, K. & Johnson-Laird, P. N. (1982) Spatial descriptions and referential continuity. *Journal of Verbal Learning and Verbal Behavior* 21:296–306. [rPNJ-L]
Ellis, W. D. (1938) *A source book of Gestalt psychology.* Harcourt, Brace. [ASL]
Erickson, J. R. (1974) A set analysis theory of behavior in formal syllogistic reasoning tasks. In: *Loyola Symposium on Cognition*, vol. 2, ed. R. Solso. Erlbaum. [aPNJ-L]
Ericsson, K. A. & Simon, H. A. (1984) *Protocol analysis: Verbal Reports as data.* MIT Press. [MDSB]
Evans, J. St.B. T. (1977) Linguistic factors in reasoning. *Quarterly Journal of Experimental Psychology* 29:297–306. [aPNJ-L]
(1987) Reasoning. *Psychological Survey* 6:74–93. [aPNJ-L]
(1989) *Bias in human reasoning: Causes and consequences.* Erlbaum. [aPNJ-L, PL]
(1991) Review of Johnson-Laird & Byrne, *Deduction.* Erlbanm. *Quarterly Journal of Experimental Psychology* 43A:916–17. [JStBTE]
(in press) Bias and rationality. In: *Rationality,* ed. K. I. Manktelow & D. E. Over. Routledge. [rPNJ-L, PE, KIM]
Evans, J. St.B. T., Barston, J. L. & Pollard, P. (1983) On the conflict between logic and belief in syllogistic reasoning. *Memory and Cognition* 11:295–306. [SEN]
Evans, J. St.B. T., Over, D. E. & Manktelow, K. I. (1992) Reasoning as decision making: One kind of rationality. *Cognition* (in press). [DO]
Falmagne, R. J. (1988) *Language as a constitutive factor in logical knowledge: Cognitive, philosophical and linguistic considerations.* Cognitive Science Laboratory, Technical Report 28, Princeton University. [RJF]
(1992) Reflections on acquisition processes. In: *Analytic approaches to human cognition,* ed. J. Alegria, D. Holender, J. Jnnca de Morais & M. Radeau. Elsevier. [RJF]
Fisher, A. (1989) *The logic of real argument.* Cambridge University Press. [AF]
Fitting, M. C. (1983) *Proof methods for modal and intuitionistic logic.* Reidel. [ADA]
Fodor, J. A. (1975) *The language of thought.* Harvard University Press. [MDSB]
(1983) *The modularity of mind.* Bradford Books/MIT Press. [NC, KMG, MO]
Ford, M. (1985) Review of "mental models." *Language* 61:897–903. [aPNJ-L]
Freeman, J. (1987) *Thinking logically.* Prentice-Hall. [AF]
Galotti, K. M. (1989) Approaches to studying formal and everyday reasoning. *Psychological Bulletin* 105:331–51. [JBaro, KMG]
Galotti, K. M., Baron, J. & Sabini, J. P. (1986) Individual differences in syllogistic reasoning: Deduction rules or mental models? *Journal of Experimental Psychology: General* 115:16–25. [arPNJ-L, KMG, RJS]
Garey, M. R. & Johnson, D. S. (1979) *Computers and intractability: A guide to the theory of NP-completeness.* Freeman. [MO]
Garnham, A. (1987) *Mental models as representations of discourse and text.* Ellis Horwood. [arPNJ-L]
(1993) Is logicist cognitive science possible? *Mind and Language* (in press). [NC, MO]
Gentner, D. & Stevens, A. L., eds. (1983) *Mental models.* Erlbaum. [MDSB]
Gigerenzer, G. & Hug, K. (1992) Domain-specific reasoning: Social contracts, cheating, and perspective change. *Cognition* 43:127–71. [DWG]
Gilbert, D. T. (1991) How mental systems believe. *American Psychologist* 46:107–19. [DWG]
Gillan, D. J. (1981) Reasoning in the chimpanzee: II. Transitive inference. *Journal of Experimental Psychology* 7:150–64. [HD]
Giordano, F. R. & Weir, M. D. (1985) *A first course in mathematical modeling.* Addison-Wesley. [ASL]
Givan, R., McAllester, D. & Shalaby, S. (1991) Natural language based inference procedures applied to Schubert's steamroller. *Proceedings of the Ninth National Conference on Artificial Intelligence.* AAAI Press/MIT Press. [JMC]
Glymour, C. (1987) Android epistemology and the frame problem: Comments on Dennett's Cognitive wheels. In: *The robot's dilemma: The frame problem in artificial intelligence,* ed. Z. W. Pylyshyn. Ablex. [MO]
Goldman, A. 1. (1986) *Epistemology and cognition.* Harvard University press. [aPNJ-L]
Goodman, N. (1966) *Fact, fiction and forecast.* Hackett. [PE]
(1968) *Languages of art.* Bobbs-Merrill. [KB]
(1954/1983) *Fact, fiction and forecast,* 4th ed. Harvard University Press. [MO]
Govier, T. (1981) *A practical study of argument* (2nd ed.). Wadsworth. [AF]
Green, D. W. (1992) Counter-examples and the selection task: Fleshing-out the model. Paper presented to the Wason Symposium at the Second International Conference on Thinking. University of Plymouth, England, July 27–30. [DWG]
Greene, S. B. (1992) Multiple explanations for multiply quantified sentences: Are multiple models necessary? *Psychological Review* 99:184–87. [PP]
Grice, H. P. (1975) Logic and conversation. In: *Syntax and semantics,* vol. 3: *Speech acts,* ed. P. Cole & J. L. Morgan. Seminar Press. [rPNJ-L]
Griffin, D. R. (1974) *Animal thinking.* Harvard University Press. [HD]
Griggs, R. A. (1983) The role of prohlem content in the selection task and in the THOG problem. In: *Thinking and reasoning: Psychological approaches,* ed. J. St.B. T. Evans. Routledge & Kegan Paul. [aPNJ-L]
Griggs, R. A. & Cox, J. R. (1982) The elusive thematic-materials effect in Wason's selection task. *British Journal of Psychology* 73:407–20. [aPNJ-L, PP]
Griggs, R. A. & Newstead, S. E. (1982) The role of problem structure in a deductive reasoning task. *Journal of Experimental Psychology: Language, Memory, and Cognition* 8:297–307. [aPNJ-L]
Gustason, W. & Ulrich, E. (1973) *Elementary symbolic logic.* Holt, Rinehart & Winston. [JHF]
Guyote, M. J. & Sternberg, R. J. (1981) A transitive-chain theory of syllogistic reasoning. *Cognitive Psychology* 13:461–525. [aPNJ-L]
Hagert, G. (1984) Modeling mental models: Experiments in cognitive modeling of spatial reasoning. In: *Advances in artificial intelligence,* ed. T. O'Shea. North-Holland. [aPNJ-L]
Hanks, S. & McDermott, D. (1985) Default reasoning, nonmonotonic logics and the frame problem. In: *Proceedings of the American Association for Artificial Intelligence.* Philadelphia, PA. [MO]
(1986) Temporal reasoning and default logics. *Computer Science Technical Report,* No. 430, Yale University. [MO]
(1987) Nonmonotonic logic and temporal projection. *Artificial Intelligence* 33:379–412. [NC]
Harman, G. (1986) *Change in view: Principles of reasoning.* Bradford Books/MIT Press. [rPNJ-L, KB]
Henle, M. (1962) The relation between logic and thinking. *Psychological Review,* 69:366–78. [ASL]
(1978) Foreword to *Human reasoning,* ed. R. Revlin & R. E. Mayer. Winston. [aPNJ-L]
Hintikka, J. (1955) Form and content in quantification theory. *Acta Philosophica Fennica* 8:7–55. [arPNJ-L, ADA]
Hoch, S. J. (1985) Counterfactual reasoning and accuracy in predicting personal events. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 11:719–31. [JBaro]
Hollis, M. (1970) Reason and ritnal. In: *Rationality,* ed. B. R. Wilson. Basil Blackwell. [aPNJ-L]
Huttenlocher, J. (1968) Constructing spatial images: A strategy in reasoning. *Psychological Review* 75:286–98. [aPNJ-L, HD]
Inder, R. (1987) *Computer simulation of syllogism solving using restricted mental models.* Doctoral Dissertation, Center for Cognitive Science, University of Edinburgh. [RI]
Inhelder, B. & Piaget, J. (1958) *The growth of logical thinking from childhood to adolescence.* Routledge & Kegan Paul. [aPNJ-L, HD, ASL]
Jackendoff, R. (1988) Exploring the form of information in the dynamic unconscious. In: *Psychodynamics and cognition,* ed. M. J. Horowitz. University of Chicago Press. [aPNJ-L]
Jackson, S. L. & Griggs, R. A. (1990) The elusive pragmatic reasoning schemas effect. *Quarterly Journal of Experimental Psychology* 42A:353–73. [PP]
Johnson-Laird, P. N. (1975) Models of deduction. In: *Reasonings: Representation and process in children and adults,* ed. R. J. Falmagne. Erlbaum. [aPNJ-L]
(1983) *Mental models: Towards a cognitive science of language, inference, and consciousness.* Cambridge University Press (Cambridge). [arPNJ-L, JBarw, BJM, SEN, MO, NEW]
(1986) Conditionals and mental models. In: *On conditionals,* ed. Ferguson, C., Reilly, J., ter Meulen, A., Tragott, E. C. Cambridge University Press. [AGBtM]
(1990a) Propositional reasoning: An algorithm for deriving parsimonious conclusions. Unpublished paper. Princeton University. [aPNJ-L]
(1990b) The development of reasoning ability. In: *Causes of development: Interdisciplinary perspectives,* ed. G. Butterworth & P. Bryant. Harvester Wheatsheaf. [rPNJ-L]
(1992) *Human and machine thinking.* Erlbaum. [rPNJ-L]
Johnson-Laird, P. N. & Bara, B. G. (1984) Syllogistic inference. *Cognition* 16:1–61. [arPNJ-L, JStBTE, DWG, WH, SEN, NEW]
Johnson-Laird, P. N. Byrne, R. M. J. (1989) *Only* reasoning. *Journal of Memory and Language* 28:313–30. [aPNJ-L]

(1990) Meta-logical problems: Knights, knaves, and Rips. *Cognition* 36:69–81. [aPNJ-L]
(1991) *Deduction*. Erlbaum.
(1992) Modal reasoning, models, and Manktelow and Over. *Cognition* 43:173–82. [rPNJ-L, KIM]
Johnson-Laird, P. N., Byrne, R. M. J. & Schaeken, W. (1992) Propositional reasoning by model. *Psychological Review* 99:418–39. [aPNJ-L]
Johnson-Laird, P. N., Byrne, R. M. J. & Tabossi, P. (1989) Reasoning by model: The case of multiple quantification. *Psychological Review* 96:658–73. [arPNJ-L]
(1992) In defense of reasoning: A reply to Greene (1992). *Psychological Review* 99:188–90. [PP]
Johnson-Laird, P. N. & Wason, P. C. (1970) Insight into a logical relation. *Quarterly Journal of Experimental Psychology* 22:49–61. [KIM]
Kamp, H. & Reyle, U. (1992) *From discourse to logic*. Kluwer Academic. [AGBtM]
Keil, F. C. (1991) The emergence of theoretical beliefs as constraints on concepts. In: *The epigenesis of mind: Essays on biology and cognition*, ed. S. Carey & R. Gelman. Erlbaum. [rPNJ-L]
Köhler, W. (1938) *The place of value in a world of fact*. Liveright. [ASL]
Koriat, A., Lichtenstein, S. & Fischhoff, B. (1980) Reasons for confidence. *Journal of Experimental Psychology: Human Learning and Memory* 6:107–18. [JBaro]
Kuhn, D. (1991) *The skills of argument*. Cambridge University Press. [JBaro].
Larkin, J. H., McDermott, J., Simon, D. & Simon, H. A. (1980) Expert and novice performance in solving physics problems. *Science* 208:1335–42. [RJS]
Lee, G. & Oakhill, J. (1984) The effects of externalization on syllogistic reasoning. *Quarterly Journal of Experimental Psychology* 36A:519–30. [DWG]
Legrenzi, P. (1970) Relations between language and reasoning about deductive rules. In; *Advances in psycholinguistics*, ed. G. B. Flores D'Arcais & W. J. M. Levelt. North-Holland. [aPNJ-L]
Legrenzi, P., Girotto, V. & Johnson-Laird, P. N. (1992) Focussing in reasoning and decision making. Manuscript submitted. [rPNJ-L]
Leslie, A. & Thaiss, L. (1992) Domain specificity in conceptual development: Neuropsychological evidence from autism. *Cognition* 43:225–51. [DWG]
Luchins, A. S. & Luchins, E. H. (1964) On awareness and denotation of contradictions. *Journal of General Psychology* 71:233–46. [ASL]
(1965a) Reactions to inconsistencies: Phenomenal versus logical contradictions. *Journal of General Psychology* 73:47–65. [ASL]
(1965b) *Logical foundations of mathematics for behavioral scientists*. Holt, Reinhart and Winston. [ASL]
(1970) *Wertheimer's seminars revisited: Problem solving and thinking*, vols. 1 & 2. Faculty Student Association, SUNY-Albany. [ASL]
(1978) *Revisiting Wertheimer's seminars*, vol. 1. *Values, social influences, and power*. Associated University Presses. [ASL]
(1987) Max Wertheimer in America: Part I. *Gestalt Theory* 9(2):70–101. [ASL]
(1991) *Max Wertheimer: His life and work*, vols. 1 & 2. Rensselaer Polytechnic Institute. [ASL]
Macnamara, J. (1986) *A border dispute: The place of logic in psychology*. Bradford Books/MIT Press. [aPNJ-L]
Madruga, J. A. G. (1984) Procesos de error en el razonamiento silogistico: Doble procesamiento y estrategia de verificacion por. In: *Lecturas de psicologia del pensamiento*, ed. M. Carretero & J. A. G. Madruga. Alianza. [aPNJ-L]
Manktelow, K. I. & Over, D. E. (1987) Reasoning and rationality. *Mind and Language* 2:199–219. [aPNJ-L]
(1991) Social roles and utilities in reasoning with deontic conditionals. *Cognition* 39:85–105. [KIM]
(1992a) Utility and deontic reasoning: Some comments on Johnson-Laird and Byrne. *Cognition* 43:183–88. [KIM]
(1992b) Rationality, utility, and deontic reasoning. In: *Rationality*, ed. K. I. Manktelow & D. E. Over. Routledge (in press). [DO]
Marr, D. (1982) *Vision: A computational investigation into the human representation and processing of visual information*. Freeman. [arPNJ-L, KB]
McCarthy, J. M. (1980) Circumscription – a form of non-monotonic reasoning. *Artificial Intelligence* 13:27–39. [NC]
McCarthy, J. M. & Hayes, P. J. (1969) Some philosophical problems from the standpoint of artificial intelligence. In: *Machine intelligence 4*, ed. B. Meltzer & D. Michie. Edinburgh University Press. [NC]
McCluskey, E. J. (1956) Minimization of Boolean functions. *Bell System Technical Journal* 35:1417–44. [aPNJ-L]
McDermott, D. (1986) *A critique of pure reason*. Technical Report, Department of Computer Science, Yale University, June, 1986. [MO]
(1987) A critique of pure reason. *Computational Intelligence* 3:151–60. [aPNJ-L, NC]
McGinn, C. (1989) *Mental content*. Basil Blackwell. [aPNJ-L]
McGonigle, B. O. & Chalmers, M. (1977) Are monkeys logical? *Nature* 267:694–96. [HD]
ter Meulen, A. (1992) *Representing time in natural language. The dynamic interpretation of tense and aspect*. Bradford Books/MIT Press (in press). [AGBtM]
Miller, G. A. & Johnson-Laird, P. N. (1976) *Language and perception*. Cambridge University Press (Cambridge). [aPNJ-L]
Morton, A. (1988) Making arguments explicit: The theoretical interest of practical difficulties. In: *Critical thinking: Proceedings of the First British Conference on Informal Logic and Critical Thinking*, ed. A. Fisher. University of East Anglia Press. [rPNJ-L]
Newell, A. (1981) Reasoning, problem solving and decision processes: The problem space as a fundamental category. In: *Attention and performance*, vol. 8, ed. R. Nickerson. Erlbaum. [aPNJ-L]
(1990) *Unified theories of cognition*. Harvard University Press. [arPNJ-L, NEW]
Newstead, S. E. (1989) Interpretational errors in syllogistic reasoning. *Journal of Memory and Language* 28:78–91. [PP]
Newstead, S. E., Manktelow, K. I. & Evans, J. St.B. T. (1982) The role of imagery in the representation of linear orderings. *Current Psychological Research* 2:21–32. [rPNJ-L]
Newstead, S. E., Pollard, P., Evans, J. St.B. T. & Allen, J. L. (1992) The source of belief bias effects in syllogistic reasoning. *Cognition*. [SEN]
Nisbett, R. E. & Ross, L. (1980) *Human inference: Strategies and shortcomings of social judgment*. Prentice Hall. [aPNJ-L]
Oakhill, J. V. & Johnson-Laird, P. N. (1985a) Rationality, memory and the search for counterexamples. *Cognition* 20:79–94. [rPNJ-L]
(1985b) The effects of belief on the spontaneous production of syllogistic conclusions. *Quarterly Journal of Experimental Psychology* 37A:553–69. [aPNJ-L]
Oakhill, J. V., Johnson-Laird, P. N. & Garnham, A. (1989) Believability and syllogistic reasoning. *Cognition* 31:117–40. [aPNJ-L, SEN]
Oaksford, M. & Chater, N. (1991) Against logicist cognitive science. *Mind and Language* 6:1–38. [NC, MO]
(1992a) Reasoning theories and bounded rationality. In: *Rationality*, ed. K. Manktelow & D. Over. Routledge. [MO]
(1992b) Bounded rationality in taking risks and drawing inferences. *Theory & Psychology* 2:225–30. [MO]
O'Brien, D. P. (in press) Mental logic and irrationality: We can put a man on the moon, so why can't we solve those logical reasoning problems? In: *Rationality*, ed. K. I. Manktelow & D. E. Over. Routledge. [MDSB, DO]
O'Brien, F. (1939/1967) *At swim-two-birds*. Penguin Books. [aPNJ-L]
Oden, G. C. (1987) Concept knowledge and thought. *Annual Review of Psychology* 38:203–27. [aPNJ-L]
Osherson, D. (1975) Logic and models of logical thinking. In: *Reasoning: Representation and process in children and adults*, ed. R. J. Falmagne. Erlbaum. [aPNJ-L]
Partee, B., ter Meulen, A. & Wall, R. (1990) *Mathematical methods in linguistics*. Kluwer. [AGBtM]
Peakcocke, C. (1986) Explanation in computation psychology: Language, perception, and level 1.5. *Mind & Language* 1:101–23. [KB]
Perkins, D. N., Faraday, M. & Bushey, B. (1991) Everyday reasoning and the roots of intelligence. In: *Informal reasoning and education*, ed. J. F. Voss, D. N. Perkins & J. W. Segal. Erlbaum. [JBaro]
Piaget, J. (1970) *Genetic epistemology* (translated by Elinor Duckworth). Columbia University Press. [HD]
(1977) *The essential Piaget: An interpretative reference and guide*, ed. H. E. Gruber & J. J. Voneche. Basic Books. [BGB]
Politzer, G. & Braine, M. D. S. (1991) Responses to inconsistent premisses cannot count as suppression of valid inferences. *Cognition* 38:103–8. [rPNJ-L, SF, DO]
Polk, T. A. & Newell, A. (1988) Modeling human syllogistic reasoning in Soar. In: *Tenth Annual Conference of the Cognitive Science Society* 181–87. Erlbaum. [aPNJ-L]
(1992) A verbal reasoning theory for categorical syllogisms. Unpublished paper. Department of Psychology, Carnegie Mellon University. [rPNJ-L, TAP]
Pollock, J. (1989) *How to build a person: A prolegomenon*. Bradford Books/MIT Press. [aPNJ-L]
Popper, K. R. (1934/1959) *The logic of scientific discovery*. Basic Books. [BGB]
Prier, R. A. (1976) *Archaic logic: Symbol and structure in Heraclitus, Parmenides, and Empedocles*. Mouton. [BJM]
Pullman, G. K. (1991) *The great Eskimo vocabulary hoax and other irreverent essays on the study of language*. University of Chicago Press. [KMG]

Pylyshyn, Z. (1980) Computation and cognition: Issues in the foundations of cognitive science. *Behavioral and Brain Sciences* 3:111–32. [RI]

(1981) The imagery debate: Analogue media versus tacit knowledge. In: *Imagery*, ed. N. Block. MIT Press. [aPNJ-L]

Quine, W. V. O. (1953) Two dogmas of empiricism. In: *From a logical point of view*, ed. W. V. O. Quine. Harvard University Press. [MO]

(1955) A way to simplify truth functions. *American Mathematical Monthly* 59:521–31. [aPNJ-L]

Rawls, J. (1970) *A theory of justice*. Harvard University Press. [PE]

Reiter, R. (1973) A semantically guided deductive system for automatic theorem-proving. *Proceedings of Third International Joint Conference on Artificial Intelligence* 41–46. [aPNJ-L]

(1978/1985) On reasoning by defanlt. In: *Readings in knowledge representation*, ed. R. Brachman & H. Levesque. Morgan Kaufman. [MO]

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

Riley, C. A. & Trabasso, T. (1974) Comparatives, logical structures, and encoding in a transitive inference task. *Journal of Experimental Child Psychology* 17:187–203. [HD]

Rips, L. J. (1983) Cognitive processes in propositional reasoning. *Psychological Review* 90:38–71. [aPNJ-L, KMG]

(1986) Mental muddles. In: *The representation of knowledge and belief*, ed. M. Brand & R. M. Harnish. University of Arizona Press. [aPNJ-L, KB]

(1989) The psychology of knights and knaves. *Cognition* 31:85–116. [aPNJ-L]

Rogoff, B. & Lave, J., eds. (1984) *Everyday cognition: Its development in social context*. Harvard University Press. [KMG]

Rumelhart, D. E. (1979) *Analogical processes and procedural representation* (Technical Report No. 81). Center for Human Information Processing, University of California, San Diego. [PP]

(1980) Schemata: The building blocks of cognition. In: *Theoretical issues in reading comprehension*, ed. R. J. Spiro, B. C. Bruce & W. F. Brewer. Erlbaum. [PP]

(1989) Towards a microstructural account of human reasoning. In: *Similarity and analogical reasoning*, ed. S. Vosniadou & A. Ortony. Cambridge University Press (Cambridge). [aPNJ-L]

Rumelhart, D., Smolensky, P., McClelland, J. L. & Hinton, G. E. (1986) Schemata and sequential thought processes in PDP models. In: *Parallel distributed processing: Explorations in the microstructure of cognition*, vol. 2: *Psychological and biological models*, ed. J. M. McClelland & D. Rumelhart. MIT Press. [JStBTE]

Russell, J. (1987) Rule-following, mental models, and the developmental view. In: *Meaning and the growth of understanding: Wittgenstein's significance for developmental psychology*, ed. M. Chapman & R. A. Dixon. Springer. [aPNJ-L]

Scribner, S. (1986) Some characteristics of practical thought. In: *Practical intelligence*, ed. R. J. Sternberg & R. Wagner. Cambridge University Press. [KMG]

Selz, O. (1935) Versuche znr Hebnng des Intelligenzniveaus: Ein Beitrag zur Theorie der Intelligenz und ihrer erziehlichen Beeinflussung. *Zeitschrift für Psychologie* 134:236–301. [JBaro]

Shallice, T. (1988) *From neuropsychology to mental structure*. Cambridge University Press (Cambridge). [DWG]

Shastri, L. & Ajjanagadde, V. G. (1990) An optimal limited inference system. *Proceedings of the Eighth National Conference on Artificial Intelligence*. AAAI Press/MIT Press. [JMC]

Shaw, G. B. (1933) The adventures of the black girl in her search for God. Cited by Skinner, B. F. (1953) *Science and human behavior*. Macmillan. [HD]

Shoham, Y. (1987) *Reasoning about change: Time and causation from the standpoint of artificial intelligence*. MIT Press. [NC]

Simon, H. A. (1991) *Models of my life*. Basic Books. [rPNJ-L]

Smullyan, R. M. (1968) *First order logic*. Springer-Verlag. [aPNJ-L, ADA]

Sperber, D. & Wilson, D. (1986) *Relevance: Communication and cognition*. Basil Blackwell. [aPNJ-L]

Stenning, K. (1992) Distinguishing conceptual and empirical issues about mental models. In: *Models in the mind*, ed. Y. Rogers, A. Rutherford & P. Bibby. Academic Press. [KS]

Stenning, K. & Oaksford, M. R. (in press) Rational reasoning and human implementations of logic. In: *Rationality*, ed. K. Manktelow & D. Over. Routledge and Kegan Panl. [KS]

Stenning, K. & Oberlander, J. (Submitted) A cognitive theory of graphical and linguistic reasoning: Logic and implementation. Research Paper 20. Human Communication Research Centre. University of Edinburgh, Edinburgh. [KS]

Stich, S. (1990) *The fragmentation of reason*. MIT Press. [PE]

Thomas, S. (1986) *Practical reasoning in natural language* (3rd ed.). Prentice Hall. [AF]

Toulmin, S. E. (1958) *The uses of argument*. Cambridge University Press (Cambridge). [rPNJ-L]

Traugott, E., ter Meulen, A., Reilly, J. & Ferguson, C., eds. (1986) *On conditionals*. Cambridge University Press. [AGBtM]

Tversky, A. & Kahneman, D. (1973) Availability: A heuristic for judging frequency and probability. *Cognitive Psychology* 5:207–32. [rPNJ-L]

Tweney, R. D. (1989) A framework for the cognitive psychology of science. In: *Psychology of science: Contributions to metascience*, ed. B. Gholson, W. R. Shadish, Jr., R. A. Neimeyer & A. C. Houts. Cambridge University Press (Cambridge). [RDT]

Tweney, R. D., Doherty, M. E. & Mynatt, C. R., eds. (1981) *On scientific thinking*. Columbia University Press. [RDT]

von Fersen, L., Wynne, C. D. L., Delius, J. & Stadden, J. E. R. (1991) Transitive inference formation in pigeons. *Journal of Experimental Psychology: Animal Behavior Processes* 17:334–41. [HD]

Wason, P. C. (1965) The context of plausible denial. *Journal of Verbal Learning and Verbal Behavior* 4:7–11. [aPNJ-L]

(1966) Reasoning. In: *New horizons in psychology*, ed. B. M. Foss. Pengnin. [arPNJ-L, PL]

(1983) Realism and rationality in the selection task. In: *Thinking and reasoning: Psychological approaches*, ed. J. St.B. T. Evans. Rontledge & Kegan Paul. [aPNJ-L]

Wason, P. C. & Green, D. W. (1984) Reasoning and mental representation. *Quarterly Journal of Experimental Psychology* 36A:597–610. [aPNJ-L]

Wason, P. C. & Johnson-Laird, P. N. (1972) *Psychology of reasoning: Structure and content*. Harvard University Press. [aPNJ-L, BGB]

Wertheimer, M. (1912) Numbers and numerical concepts in primitive peoples (abridged translation). In: *A source book of Gestalt psychology*, ed. W. D. Ellis (1938). Harcourt, Brace. [ASL]

(1925) The syllogism and prodnctive thinking (abridged translation). In: *A source book of Gestalt psychology*, ed. W. D. Ellis (1938). Harcourt, Brace. [ASL]

(1934) On trutb. *Social Research* 1:135–46. [ASL]

(1945) *Productive thinking*. Enlarged edition (1959). Harper. [ASL]

(1961) *Productive thinking*. Tavistock. [JStBTE]

Wetherick, N. E. (1989) Psychology and syllogistic reasoning. *Philosophical Psychology* 2:111–24. [NEW]

(1991) What goes on in the mind when we solve syllogisms? In: *Mental images in human cognition*, ed. R. H. Logie & M. Denis. North Holland. [NEW]

(in press) Human rationality. In: *Rationality*, ed. K. I. Manktelow & D. E. Over. Routledge. [NEW]

Wetherick, N. & Gilhooly, K. (1990) Syllogistic reasoning: Effects of premise order. In: *Lines of thinking: Reflections on the psychology of thought*, vol. 1, ed. K. J. Gilhooly, M. T. G. Keane, R. H. Logie & G. Erdos. Wiley. [aPNJ-L, NEW]

Wise, M. N. (1979) The mutual embrace of electricity and magnetism. *Science* 203:1310–18. [rPNJ-L]

Wood, D. J. (1969) The nature and development of problem solving strategies. Unpublished Doctoral Dissertation. University of Nottingham. [RJS]

Woodworth, R. S. (1938) *Experimental psychology*. Holt. [MDSB]

Woodworth, R. S. & Sells, S. B. (1935) An atmosphere effect in formal syllogistic reasoning. *Journal of Experimental Psychology* 18:451–60. [aPNJ-L]

Yates, J. (1985) The content of awareness is a model of the world. *Psychological Review* 92:249–85. [rPNJ-L]

Yule, P. & Stenning, K. (1992) The figural effect and a graphical algorithm for syllogistic reasoning. In: *Proceedings of the Fourteenth Annual Conference of the Cognitive Science Society*, Bloomington, Indiana. [KS]