MENTAL MODELS AND DEDUCTIVE REASONING

By

P.N. Johnson-Laird

Final version

September 14th 2005

To be published in Rips, L. and Adler. J. (Eds.):

Reasoning: Studies in Human Inference and Its Foundations

Cambridge: Cambridge University Press

Number of words (including references): 11,467

Author's address

Department of Psychology

Princeton University

Green Hall

Princeton

NJ 08544

USA

tel: (609) 258 4432

fax: (609) 258 1113

email: phil@princeton.edu

Key words for subject index

Conditionals

Conjunction

Counterexamples

Deduction

Disjunctions

Iconicity

Illusory inferences

Inconsistency

Logic

- Formal rules of inference
- Truth-functional connectives

Mental models

Negation

Possibilities

Reasoning:

- Deductive
- Illusions in
- Inductive
- Relational
- Role of counterexamples
- Strategies in

Semantics

- Formal semantics

Truth

Validity

How do people reason? The view that I learned at my mother's knee was that they rely on logic. During the 1960's and 1970's when the study of thinking had become respectable again after the Dark Ages of Behaviorism, psychologists – including the present author – took this view for granted. The idea that logic provided the norms of reasoning can be traced back to the rise of modern logic, and was defended in the nineteenth century by both Boole (1854) and Mill (1874). In the twentieth century, the Swiss psychologist, Jean Piaget, and his colleagues argued that the construction of a formal logic in the mind was the last great step in children's intellectual development, and that it occurred at about the age of twelve (see, e.g., Inhelder and Piaget, 1958). And so thirty years ago the task for psychologists appeared to be to determine which particular formal logic was laid down in the mind and which particular rules of inference were used in its mental formulation. That, at least, was how several like-minded authors conceived their research (see, e.g., Osherson, 1974-6; Johnson-Laird, 1975; Braine, 1978; Rips, 1983). In the parallel 'universe' of artificial intelligence, researchers were similarly developing computer programs that proved theorems relying on formal rules of inference (e.g., Bledsoe, 1977). The main skeptics were those engaged in trying to analyze everyday arguments. They discovered that it was extraordinarily difficult to translate such arguments into formal logic. As a result, many of them abandoned logic as a method of analysis (see, e.g., Toulmin, 1958; Scriven, 1976), and later they went on to found a pedagogical society, AILACT (Association for Informal Logic and Critical Thinking), in which they advocated a variety of other methods of analysis.

The event that woke me from my dogmatic slumbers was the late Peter Wason's discovery of the effects of content on his "selection" task. In the abstract version of the

task (Wason, 1966), the experimenter lays down four cards in front of you, such as: A, B, 2, 3. You know that each card has a number on one side and a letter on the other side. Your task is to select just those cards to turn over that are relevant to determining whether a general claim is true or false. The general claim in one version of the task is:

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

Like all the studies described in the present article, the experiment tested "naïve" individuals – I use the term, not to impugn their intelligence, but merely to mean that they had received no training in logic. In the selection task, they tended to choose the A card, and sometimes the 2 card. But, they rarely chose the 3 card. You can grasp the need to do so if you think about the consequences of turning over the A card. If a 3 is on its other side, the general claim is false. By parity of argument, if you turn over the 3 card and find an A on its other side, the general claim is also false. Hence, you do need to select the 3 card. If you select the 2 card, then nothing on its other side can show that the general claim is false, unless you take the claim to mean: if and *only if* a card has an 'A' on one side then it has a '2' on its other side. In this case, however, you ought to select all four cards.

The selection task was inspired by Popper's philosophy of science. He had argued that what divides science from non-science is "falsifiability": a scientific hypothesis is one that in principle observations could show to be false (Popper, 1959). Any hypothesis that could not be refuted in this way is outside the borders of science. What Wason discovered is that individuals unfamiliar with logic or with Popper's idea have a perverse propensity *not* to try to falsify general claims. The failure was embarrassing to Piagetians, because Piaget had argued that once children attain the level

of "formal operations" – the level corresponding to the acquisition of logic, they would check the truth of conditionals of the form, *If A then B*, by searching for counterexamples of the form: *A and not-B*. But, the participants signally failed to reason in this way: the 3 card corresponds to not-B, and they did not select it. For anyone who had studied logic, this error of omission was puzzling. And so Wason and I embarked on a three-year research program, helped by a graduate, Diana Shapiro, to try to understand what was going on.

Why don't individuals select that 3 card? There were many possibilities. The experimental procedure might somehow mislead them. Psychological experiments are social interactions in a microcosm, and sensitive to all sorts of unforeseen factors. The participants may construe the task in ways quite different from what the experimenter has in mind. The psychologist may use inappropriate materials or procedures, or fail to control the experiment properly. The instructions explain the task, but the participants also think for themselves. This point became vivid to me when I read a participant the instructions, and he then said, "Yes, but what do you *really* want me to do?" We tried all sorts of experimental manipulations in order to elicit a correct performance. We changed the form of the rule; we put all the information on one side of a card; we gave participants brief 'intellectual psychotherapy'. None of these manipulations did much good (see Wason and Johnson-Laird, 1972).

Wason then suggested changing the *content* of the cards and the claim. Perhaps because my head was still stuffed with formal logic, I thought that this manipulation was crazy (though I didn't say so). Wason and Shapiro went ahead with the experiment. They used a general claim with an underlying conditional meaning:

Every time I go to Manchester I travel by train

and four cards about journeys, with a destination on one side and a mode of transport on the other side: Manchester, Leeds, train, car. The participants were now much more likely to realize that a journey by *car* (the *not-B* case) was pertinent to the truth or falsity of the claim, and to select this card. If the destination on the other side was Manchester, then the claim was false (Wason and Shapiro, 1971). It is hard to convey the stunning nature of this result at the time. A change in content alone had a striking effect on reasoning, even though the two sorts of contents were identical in formal structure. Paolo and Maria Legrenzi and I carried out a similar experiment using a regulation about the postage required on sealed envelopes. The same robust change in performance occurred, and it did not transfer to the abstract version of the task (Johnson-Laird, Legrenzi, and Sonino Legrenzi, 1972).

The true reason for the difference in performance between abstract and concrete selection tasks remains a matter of controversy, and several hypotheses purport to explain it (e.g., Cheng and Holyoak, 1985; Cosmides, 1989; Johnson-Laird and Byrne, 1991; Oaksford and Chater, 1998; see also the chapters in this volume by Fodor and by Sperber). But, at the time, the phenomenon convinced several psychologists that formal logic could not explain human reasoning. Formal logic, by definition, is concerned solely with the logical form of assertions, not their content. Yet, the selection task showed that content mattered just as much as form in reasoning (Wason and Johnson-Laird, 1972).

The search for an alternative theory took some time, but it led to the topic of the present chapter, the theory of mental models or the "model theory" for short. The

chapter begins with an outline of this theory. It examines in some detail the theory's account of negation, and of how one model is conjoined with another. Unlike logic, the interpretation of sentences in daily life is often modulated by knowledge, both a knowledge of what is being referred to and general knowledge. The chapter explains these phenomena, and their consequences on reasoning. It considers how individuals develop strategies for reasoning with particular sorts of problem. And, finally, it draws some general morals about reasoning.

The mental model theory

The mental model theory has various antecedents of which the most important is the use of models in logic to provide a systematic method for the interpretation of assertions in logical calculi. However, this article is concerned, not with the history of the theory (for that, see Johnson-Laird, 2004), but with its current formulation. This section accordingly outlines its main principles. Its fundamental assumption is that reasoning is about possibilities. When we read a description of a situation, we try to envisage the alternative possibilities with which the description is compatible. If someone tell us:

My house is in the middle of the street,

we construct a mental model of a single possibility. The proposition could be true in infinitely many ways (as model theory in logic recognizes), but we cannot hold in mind such an infinitude. So, strictly speaking, our mental model of the situation captures what is common to this swarm of possibilities, namely, that the speaker's house is (roughly) in

the middle of the street rather than towards one end or the other. The sort of diagram that we might draw to convey this proposition is along the following lines:

House House House Speaker's house House House House

It has an interesting property, which the great nineteenth century philosopher and logician, Charles Sanders Peirce, called "iconicity" (see, e.g., Peirce, 1931-1958, Vol. 4, paragraph 433). He meant that the structure of such a diagram is the same as the structure of what it represents, and so the parts of the diagram map onto the parts of the scene, and the relations among the parts of the diagram are the same as the relations among the parts of the scene. Mental models are similarly as iconic as possible. An iconic representation, as Peirce (Vol. 2, paragraph, 279, and Vol. 4, paragraph 530) pointed out, has the advantage that it can yield a conclusion that was not asserted by any of the premises used in its construction. As a simple example, consider the following premises:

The cup is on the right of the saucer.

The spoon is on the left of the saucer.

An iconic diagram of the possibility compatible with these premises is as follows:

spoon saucer cup

where the left-to-right axis in the diagram corresponds to the left-to-right axis of the scene. The diagram shows that the cup is on the right of the spoon, and this conclusion follows from the premises, but it was not asserted in them. The inferential system must work according to the principle that in such cases the axes have a spatial interpretation. Various ways exist to keep track of this information, but no-one knows which way the human system uses. In contrast, the use of formal logic calls for a single non-iconic

notation for all inferences, additional premises about the logical properties of the spatial terms, and the use of formal rules of inference.

We have difficulty in thinking about more than one possibility at a time. Working memory, which holds models in mind whilst we cogitate, is limited in its capacity (see, e.g., Baddeley, 1986). Hence, we much prefer to consider just a single possibility at a time, and a sure way to get us to make mistakes is to overload us with alternative possibilities. The word "or" in English is an excellent device for causing us problems, as several experiments have shown (Bauer and Johnson-Laird, 1993). The logical meaning of "or" in an assertion, such as:

The broadcast is on network TV or it's on the radio, or both

conveys three different possibilities, and individuals can readily list them, as I have done here in abbreviated form on separate horizontal lines:

network TV

radio

network TV radio

Reasoning becomes very difficult if we have to juggle such possibilities in our minds. Hence, if, in addition, we receive a second premise:

The broadcast isn't on the radio or it's on cable TV, or both.

we have to consider the different possibilities compatible with both the first premise and this new one. That's difficult, and few of us can cope with the task.

One way to carry out the task of multiplying possibilities – a method used in a program that I wrote – is, first, to flesh out the possibilities to make them fully explicit. The two premises are:

1. The broadcast is on network TV or it's on the radio, or both.

2. The broadcast isn't on the radio or it's on cable TV, or both.

These premises are compatible with the following two sets of possibilities respectively:

1. network TV not radio

not network TV radio

network TV radio

and:

2. not radio not cable TV

radio cable TV

not radio cable TV

Two simple procedures multiply the two sets of possibilities in order to yield those that are compatible with both premises. The first procedure is used when one possibility, such as: *not on radio*, is incompatible with another, such as: *on radio* (see procedure 2 in Table 1 below). Their conjunction yields the null model, which is a special model, akin to the empty set, that represents impossibilities. The second procedure is that if one possibility is consistent with another, their conjunction combines the information in both of them without redundancy (see procedure 3 in Table 1).

The two procedures can be applied to the two sets of models above. The three possibilities for the first premise have to be multiplied by three possibilities for the second premise, i.e., a total of nine conjunctions. The conjunction of the first possibility for the first premise with the first possibility for the second premise:

network TV not radio

and:

not radio not cable TV

is consistent, and yields the following possibility:

network TV not radio not cable TV

The conjunction of the first possibility for the first premise with the second possibility for the second premise is inconsistent – one asserts that the broadcast wasn't on radio and the other assertions that it was – and so the conjunction yields the null model. The nine conjunctions yield the following possibilities, where I have omitted the null models because they represent impossibilities:

network TV not radio not cable TV

network TV not radio cable TV

not network TV radio cable TV

network TV radio cable TV

These possibilities support various conclusions, notably:

The broadcast is on network TV or it's on cable TV, or both.

This conclusion holds for all the possibilities compatible with the premises, and so, as logicians say, it is *valid*, i.e., it must be true given that the premises are true.

Naïve individuals are most unlikely to make inferences in this way. The method is a good one, but it is too demanding on working memory. One assumption of the model theory is accordingly the principle of *truth*: mental models represent clauses in the premises, affirmative or negative, only when they are true, and not when they are false (Johnson-Laird and Savary, 1999). As an example, consider again the disjunction:

The broadcast isn't on the radio or it's on cable TV.

Its three fully explicit models are shown above. But, its mental models depend on the principle of truth, and so each of its two clauses (the broadcast isn't on radio, it's on cable TV) is represented in a possibility only when it is true. Hence, the disjunction has the mental models:

not radio

cable TV

not radio cable TV

In the model of the first possibility, for example, the second clause of the disjunction is false, and so it isn't represented in the model.

Readers sometimes assume that mental models merely represent whatever clauses occur in assertions, and do not represent what is not explicitly mentioned in assertions. But, if that were so, then assertions containing the two clauses above would have the same mental models regardless of whether the sentential connective interrelating them was "if", "and", or "or". The right way to think of the principle of truth is therefore that mental models represent only those states of affairs that are possible given an assertion, and that within each of these possibilities they represent a clause in the assertion, whether it is affirmative or negative, only if it is true in that possibility. (Table 1 below spells out explicitly just how such a system can work.) Individuals may make a mental footnote about what is false in a possibility, and, if they retain it, then they can try to flesh out mental models into fully explicit models. Mental models lighten the load on working memory, because they represent less information. But, as we'll see, they have some unexpected and unfortunate consequences.

Even with mental models, we may not try to multiply out the possibilities. The model theory allows that we can develop a variety of strategies for reasoning, which I describe presently. One strategy is to make a supposition – for the sake of inference, we assume that a particular proposition holds. The previous premises are:

The broadcast is on network TV or it's on the radio, or both.

The broadcast isn't on the radio or it's on cable TV, or both.

Hence, we might think to ourselves: suppose that the broadcast isn't on network TV. We can try to combine a model of this possibility with the first mental model of the first premise, but there's an inconsistency. We can combine it with the second model, which yields the possibility that the broadcast is on the radio. So, if the broadcast isn't on network TV, then it's on the radio. We can combine the model of this possibility with the models of the second premise, and it is compatible with only one of them: the broadcast is on cable TV. It follows that if the broadcast isn't on network TV, then it's on cable TV. A moment's thought should convince you that this claim is equivalent to the disjunctive conclusion, which I proved earlier: The broadcast is on network TV or it's on cable TV.

One problematical connective is the conditional: "if". An assertion such as:

If the broadcast isn't on network TV, then it's on cable TV

is compatible with the possibility:

not network TV cable TV

But, suppose that the broadcast *is* on network TV, what does the conditional assert then? The answer, which is borne out by psychology evidence, is that it allows that the

broadcast may, or may not, be on cable TV (see, e.g., Barrouillet, Grosset, and Leças, 2000). The conditional is accordingly compatible with three possibilities:

not network TV cable TV

network TV not cable TV

network TV cable TV

These possibilities are the same for as those for the disjunction:

The broadcast is on network TV or it's on cable TV, or both.

This fact explains the ability of individuals to paraphrase conditionals as disjunctions and vice versa (Ormerod and Richardson, 2003).

When individuals reason from conditionals, they normal rely on mental models, and the theory postulates that the mental models of conditionals represent the initial possibility but use a wholly implicit model (shown as an ellipsis here) to represent the possibilities in which the antecedent clause (the if-clause) is false:

not network TV cable TV.

. . .

One consequence of these models is that inferences of the sort known as "modus ponens" are easy:

If the broadcast isn't on network TV, then it's on cable TV.

The broadcast isn't on network TV.

Therefore, it's on cable TV.

In contrast, inferences of the sort known as "modus tollens" are difficult:

If the broadcast isn't on network TV, then it's on cable TV.

The broadcast isn't on cable TV.

Therefore, it's on network TV.

To make this inference, reasoners have either to flesh out their mental models into fully explicit models, or to use some other strategy such as the suppositional one.

An inference can fail to be valid in two different ways. One way is for its conclusion to be inconsistent with the premises. Consider this putative inference:

The broadcast is on network TV or it's on the radio, or both.

The broadcast isn't on the radio or it's on cable TV, or both.

Therefore, the broadcast is neither on network TV nor on cable TV.

This conclusion corresponds to the possibility:

not network TV not cable TV

And this possibility is inconsistent with each of the four possibilities compatible with the two premises (shown earlier). Hence, the conclusion, far from following from the premises, contradicts them.

Another way in which an inference can be invalid is if its conclusion fails to follow from the premises, though it is consistent with them. Consider this putative conclusion from the previous premises:

The broadcast is on network TV and it is on cable TV.

This conclusion is compatible with the premises: it corresponds to the second and fourth possibilities listed above. So, how can we discover that it is invalid? If we use formal rules of inference, one method is to show that no proof yields the conclusion from the premises. Such a method would be exhausting, and would predict that the task of establishing invalidity would be interminable. Reasoning with mental models, in contrast, allows us a much simpler method. We merely find a *counterexample* to the

conclusion, where a counterexample is a possibility consistent with the premises but inconsistent with the conclusion (Johnson-Laird and Hasson, 2003). Such a counterexample to the previous conclusion is, for example:

network TV not cable TV

So much for the general principles of the model theory. In order to explain a crucial prediction, the next section details the procedures for constructing mental models.

The procedures for negation and conjunction

When individuals draw conclusions from a set of premises, they envisage possibilities compatible with the premises. They can cope with this task for two or three possibilities (García-Madruga, Moreno, Carriedo, Gutiérrez, and Johnson-Laird, 2001), but the task gets progressively harder when the number of possibilities increases beyond this point (Bauer and Johnson-Laird, 1993). But, what are the procedures for combining possibilities? The answer is that they depend on negation and conjunction.

A problem that is harder than it seems at first is to list the possibilities given the following assertion:

It is not the case that both Pat is here and Viv is here.

The task isn't trivial, because we don't have the answer in memory: we have to work it out. And how do we that? First, we have to work out what the unnegated proposition means:

Both Pat is here and Viv is here.

It allows just one possibility, which I'll abbreviate as:

Pat Viv

The negative proposition rules out this possibility to leave its complement, i.e., all the other possibilities. The first one that we're likely to think of is the mirror image of the possibility above in which neither Pat nor Viv is here:

not Pat not Viv

Some individuals go no further, but others are likely to realize that there are two other possibilities, in which one or other of the two individuals isn't here. Hence, the proposition as a whole is compatible with three possibilities:

not Pat not Viv

not Pat Viv

Pat not Viv

In general, the way to interpret a negative proposition is to take the propositions that occur within the negation, and to work out all the possible combinations of them and their negations. One removes from these combinations those that are compatible with the unnegated proposition, and the remaining possibilities are those for the negative proposition. No wonder that people do not cope very well with the task of listing the possibilities compatible with negative assertions (Barres and Johnson-Laird, 2003). They tend to be better at negating a disjunction than a conjunction, perhaps because the former yields fewer models than the latter. They often assume that the negation of a complex proposition consists solely of a conjunction of the negations of each of its constituent propositions.

A disjunction of alternative possibilities can be represented as a list of alternative models. To combine two such sets of models according to any logical relation between them, calls only for negation, which I've described, and logical conjunction. When

individuals interpret a set of premises, their task is to construct a model of an initial clause or premise, and then to update this model from the remaining information in the premises. Table 1 summarizes the procedures for conjoining one model with another, and it illustrates them for the conjunction of the premises:

If Pat is here then Viv is here.

Mo is here or else Pat is here, but not both.

Before you read Table 1, you might like to think for a moment what possibilities are compatible with the two premises. Most people think that there are two: Pat and Viv are here, or else Mo is here.

Insert Table 1 about here

The procedures in Table 1 apply to both mental models and fully explicit models. In both cases, the core of the interpretative process is the conjunction of one model with another. Although the mental models of the preceding premises yield a valid conclusion (see Table 1), they omit a possibility compatible with the premises. The conjunction of fully explicit models shows that a third possibility is:

not-Pat Viv Mo

The same conclusion as the one in Table 1 still follows, but reasoners who rely on mental models will fail to envisage this possibility. They should think that it is impossible for both Viv and Mo to be here. This prediction is typical of the model theory.

Mental models are based on the principle of truth, and, as I mentioned, they yield a crucial prediction, which I illustrate in two contrasting inferences. The first is:

Either Jane is kneeling by the fire and she is looking at the TV or else Mark is standing at the window and he is peering into the garden.

Jane is kneeling by the fire.

Does it follow that she is looking at the TV?

Most people say: "yes". The second inference has the same initial premise, but it is followed instead by the categorical assertion:

Jane is not kneeling by the fire.

and the question is:

Does it follow that Mark is standing by the window?

Again, most individuals say: "yes". Let's see what the theory predicts.

The first premise in both inferences is an exclusive disjunction of two conjunctions. The theory predicts that individuals should rely on mental models, and use the procedures in Table 1. Hence, they should use the meaning of the first conjunction, Jane is kneeling by the fire and she is looking at TV, to build a mental model representing this possibility, which I abbreviate as follows:

Jane: kneeling looking

They should build an analogous model of the second conjunction:

Mark: standing peering

These two models must now be combined according to the exclusive disjunction in the premises. An exclusive disjunction of the form, *either A or else B*, has two mental models, which represent its clauses only in the possibilities in which they are true. Hence, the exclusive disjunction as a whole has the mental models:

Jane: kneeling looking

Mark: standing peering

In the first inference, the categorical premise is:

Jane is kneeling by the fire

According to Table 1, its conjunction with the preceding models of the disjunction yields:

Jane: kneeling looking

And so individuals should respond: yes, Jane is looking at the TV. This analysis may strike you as glaringly obvious.

In fact, the inference is a fallacy. The principle of truth postulates that individuals normally represent what is true in possibilities, but not what is false. When I first wrote a computer program to simulate the theory, and inspected its output, I thought that there was an error in the program. I searched for this "bug" for half a day, before I discovered that the program was correct, and the error was in my thinking. What the program revealed is that for some inferences a discrepancy occurs between mental models and fully explicit models. The theory accordingly predicts that individuals should reason in a systematically fallacious way for these inferences. In some cases, the fallacies are so compelling that they resemble cognitive illusions, and so my colleagues and I refer to them as "illusory" inferences.

If you succumbed to the illusion, then you are in the company of Clare Walsh and myself. We worked with the inferences above for two days, in designing an experiment on another topic, before we realized that we were making an illusory inference, and that was *after* the discovery of other sorts of illusion. The fully explicit models of the disjunctive premise reveal the correct conclusion. The disjunction has six fully explicit models, because when one conjunction is true, the other conjunction is false, and you will remember that there are three possibilities compatible with the falsity of a conjunction. So, suppose that Mark is standing at the window and peering into the garden, then Jane

can be kneeling provided that she isn't looking at the TV. This possibility is a counterexample to the illusory inference.

The second problem had the categorical premise that Jane is not kneeling by the fire, and posed the question of whether it followed that Mark is standing by the window. Most people respond, "yes", which is a conclusion supported by the mental models shown above. This inference *is* valid. Walsh and I carried out an experiment examining a series of illusory inferences and control problems of this sort. The participants were much more likely to respond correctly to the control problems (78% correct) than to the illusory problems (10% correct), and all but one of participants showed this difference (Walsh and Johnson-Laird, 2004). Analogous illusions occur in many other domains – from reasoning about probabilities to reasoning about whether a set of assertions is consistent (for a review, see Johnson-Laird, 2005). Certain logical systems can be formulated so that they yield the same valid deductions whether they are based on models or on formal rules (Jeffrey, 1981). The same principle holds for psychological theories based on mental models or on formal rules (Stenning and Yule, 1997). But, because the theories differ in the procedures they use, a way to distinguish between them empirically is, for example, in the systematic errors that they predict. Illusory inferences are accordingly a crucial test for mental models, because no other current theory including those based on formal rules predicts them.

Semantic and pragmatic modulations and their effects on inference

The picture of reasoning that you are likely to have developed from the theory so far is a logical one. We use premises to construct mental models, which are sometimes

fleshed out into fully explicit models. If we reason with fully explicit models and make no mistakes, then our reasoning is logical. Unfortunately, our capacity to hold information in working memory is limited, and so we tend to reason with mental models, which represent only what is true, and which can accordingly lead us to make illusory inferences. Nevertheless, it may seem that our method of reasoning is logical, and so you might suppose that the logic and psychology of reasoning are quite similar. In fact, another gap exists between them. What causes this divergence is that when human beings reason, they take their knowledge into account. As a result, they often go beyond the explicit information given to them, and take a step into inductive reasoning (as knowledge is fallible).

Suppose that the following claim is true:

Ed played soccer or he played some game.

If you learn that Ed didn't play soccer, you can infer that at least he played some game. The procedures that I have described yield this inference. Its formal pattern is valid in logic, and at first sight it seems to be valid in life – to the point that when theorists postulate formal rules (see, e.g., Johnson-Laird, 1975; Rips, 1994), they include a rule of the form:

A or B.

Not A.

Therefore, B.

Consider the premise again:

Ed played soccer or he played some game.

And suppose that you learn:

Ed didn't play any game.

This fact denies the second clause of the premise, and so you should infer:

He played soccer.

This inference also follows from a formal rule for disjunction. Yet, the inference is absurd. No one in her right mind (apart from a logician) would draw it. The reason is obvious. You know that soccer is a game. That's part of the *meaning* of the word "soccer". So, if Ed didn't play any game, he can't have played soccer.

Your knowledge of the meaning of the word, "soccer," blocks the construction of a possibility in the interpretation of the assertion. And so the assertion is compatible with just two possibilities. In one of them, Ed played soccer; and in the other, he played a game. So, either way, he played a game, and the possibility that knowledge blocks is the one in which Ed played soccer but didn't play a game. The second premise asserts that Ed didn't play a game, and so it eliminates both possibilities. The second premise contradicts the first, and individuals tend to infer:

Ed didn't play soccer

This analysis shows that the meaning of a word can modulate the interpretation of a sentential connective.

General knowledge and knowledge of the context can similar pragmatic modulations. Consider this inference, for example:

If Pat entered the elevator then Viv got out one floor up.

Pat entered on the second floor.

Therefore, Viv got out on the third floor.

You envisage a possibility in which Pat entered the elevator on the second floor, Viv was already in it, the two of them traveled up together to the next floor up, the third floor, and then Viv got out. You infer this sequence of events from your knowledge of how elevators work. In such cases, pragmatic modulation adds information about temporal and spatial relations between the events referred to in a conditional (Johnson-Laird and Byrne, 2002).

Another effect of knowledge is to lead individuals to flesh out mental models into fully explicit models. Given the assertion:

Either the roulette wheel comes up red or else Viv is bankrupt

you are likely to envisage two possibilities. In one, the wheel does come up red and Viv isn't bankrupt; in the other the wheel doesn't come up red, and as a result Viv loses all her money and goes bankrupt:

Wheel: red Viv: not bankrupt

not red bankrupt

You are likely to represent both possibilities in a fully explicit way, unlike your interpretation of disjunctions that do not engage your knowledge, such as: "Either there is a triangle or else there is a circle, but not both". Your knowledge overrides the principle of truth, and you think about both possibilities in full.

Modulation can help or hinder reasoning. These effects have been demonstrated in studies in which the participants make the same form of inference with different contents (e.g., Johnson-Laird and Byrne, 2002). Assertions such as: *Pat is in Rio or she is in Norway*, elicit knowledge that one person cannot be in two places at the same time, and so the truth of the first clause implies the falsity of the second clause. In an

unpublished experiment carried out in collaboration with Tom Ormerod, participants were accordingly faster and more accurate in drawing conditional conclusions from such disjunctions than from those with neutral contents, such as *Pat is in Rio or Viv is in Norway*. Neutral contents in turn led to faster and to a greater number of valid inferences than contents, such as: *Pat is in Rio or she is in Brazil*, for which, contrary to the disjunctive form of the assertion, the truth of the first clause implies the truth of the second clause.

Experiments with conditionals have also corroborated the same effects on reasoning from conditional premises (Johnson-Laird and Byrne, 2002). Conditionals describing spatial inclusions enhanced modus tollens. For example, the premises:

If Bill is in Rio de Janeiro then he is in Brazil.

Bill is not in Brazil.

readily yielded the conclusion:

Bill is not in Rio de Janeiro

Reasoners knew that Rio de Janeiro is in Brazil, and so if Bill is not in Brazil then he cannot be in Rio. In contrast, spatial exclusions inhibited modus tollens. For example, given the premises:

If Bill is in Brazil then he is not in Rio de Janeiro.

Bill is in Rio de Janeiro.

reasoners tended to balk at the conclusion:

Bill is not in Brazil.

They knew that if Bill is in Rio then he must be in Brazil. The participants in the experiment drew twice as many modus tollens inferences from the spatial inclusions than from the spatial exclusions.

Most theories of reasoning allow for pragmatic effects, and the semantic and pragmatic modulation of mental models explains how these effects occur according to the model theory. It may be possible to explain them without recourse to models in, say, a framework based on formal rules of inference. However, as Byrne and I argued, the potential for meaning and knowledge to modulate the interpretation of connectives means that the system for interpreting sentences must always be on the lookout for such effects. It must always examine the meaning and reference of clauses to check whether they and the knowledge that they elicit modulate interpretation. This step must occur even for examples that turn out to receive a logical interpretation. The system for interpreting sentential connectives cannot work in the "truth functional" way that logic works, taking into account only the truth values of clauses (see, e.g., Jeffrey, 1981). It must take meaning, reference, and knowledge, into account. That is why the process of interpretation is never purely logical. The fact that modulation can add spatial and temporal relations between the events described in a sentence means that sentences of a given form, such as conditionals or disjunctions, have an indefinite number of different interpretations and cannot be interpreted as truth functions – an implication that has eluded some commentators on the model theory (pace Evans, Over, and Handley, 2005).

The development of strategies for reasoning

You might suppose that individuals are equipped with a single deterministic strategy for reasoning, which unwinds like clockwork. But, over the years, psychologists have discovered various embarrassments for this view. The order of the premises, for instance, has robust effects on inferences from conditional premises. It is easier to make a modus tollens inference when the categorical premise is presented before the conditional than vice versa – presumably the categorical can immediately block the representation of the otherwise salient case in which the antecedent is true (Girotto, Mazzocco, and Tasso, 1997). Such effects seem inconsistent with a single inferential strategy. Perhaps the principal reason for postulating a single strategy, however, is that studies were often insensitive to strategy. They used no more than two premises; they recorded only the conclusions that the participants drew and perhaps how long it took them to draw them. Such data cannot reveal how the participants reached their conclusions. In contrast, studies with three of four premises, even though the inferences were easy to make, revealed that individuals spontaneously developed a variety of strategies.

Consider, for instance, the following problem from a study of strategies (Van der Henst, Yang, and Johnson-Laird, 2002):

> There is a red marble in the box if and only if there is a brown marble in the box.
>
> Either there is a brown marble in the box or else there is a gray marble in the box, but not both.
>
> There is a gray marble in the box if and only if there is a black marble in the box.
>
> Does it follow that: If there is not a red marble in the box then there is a black marble in the box?

When most individuals encounter such a problem for the first time, they are nonplussed for a moment. Gradually, however, they work out its solution. The problem is, in fact, so easy that they almost always get it right. Over the course of a few problems of a similar sort, they develop a strategy, and different individuals develop different strategies. An obvious corollary is equally important. Individuals do not come to the psychological laboratory already armed with strategies for these sorts of inferences.

The correct answer to the preceding problem is: yes, if there isn't a red marble in the box then there is a black marble. As you think about the problem, you will see what is meant by a strategy, which my colleagues and I define as follows:

> A *strategy* in reasoning is a systematic sequence of elementary steps that an individual follows in making an inference.

When you learn long multiplication, you learn a strategy, but it is a deterministic one. Each juncture leads to just one next step. When you develop a strategy for reasoning, however, it doesn't fix the sequence of steps in such a rigid way. Each step in a strategy is a *tactic*. The procedures underlying a tactic are seldom, if ever, available to consciousness. You can't introspect on what enables you to combine two premises to draw a conclusion (see Table 1). But, individuals do report the particular tactical steps that they make when they carry out inferences such as the one above.

In the study from which the preceding example was taken (Van der Henst et al., 2002), the premises of each problem were compatible with just two possibilities, and half of them were presented with a valid conclusion and half of them were presented with an invalid conclusion. The participants had to think aloud as they reasoned, and they were allowed to use paper and pencil. A video-camera was above them and focused down on

the desk at which they sat. The participants occasionally made uninterpretable remarks, and they also made false starts that petered out. But, everyone correctly evaluated every problem, and it was clear what strategies they had used for nearly every problem. Sometimes they changed from one strategy to another in the middle of a problem. Overall in this experiment and subsequent ones, the participants developed five distinct strategies, which I will outline.

1. Integrated diagrams. This strategy relies on the construction of a single diagram that integrates all the information from the premises, very much along the lines of a set of mental models of the premises. With the problem above, a participant read the first premise aloud: "There's a red marble if and only if there's a brown marble", and made an immediate inference, "If brown then red". He then drew a simple diagram of this conclusion, where the arrow presumably denoted the conditional relation:

brown → red

He read the second premise aloud, "Brown or else gray", and added an element to his diagram to represent an alternative possibility:

gray

brown → red

His performance on earlier problems showed that he represented separate possibilities on separate lines. He read the third premise: "There's a gray marble if and only if there's a black marble", inferred: "If gray then black", and added a new referent to his diagram:

gray → black

brown → red

The diagram supports the conclusion: if not red then black, which he accepted. He then checked the inference, working through the premises again, and comparing them with his diagram of the two possibilities. His protocol allowed us to identify his strategy and its component tactics. But, he said nothing along the following lines, "my aim is to build a single diagram based on all the premises from which I can check whether the conclusion follows". And, as one would expect, he said nothing about the procedures underlying the tactical steps. Some participants who used this strategy drew a vertical line down the page and wrote down the colors of the marbles in the two possibilities on either side of it. Others, as in the preceding protocol, arranged the possibilities horizontally. One participant merely drew circles around the terms in the premises themselves to pick out one of the two possibilities. A telltale sign of the integrated diagram strategy is that the participants work through the premises in the order in which they are stated, and they include in their diagrams information from premises that are irrelevant to evaluating the conclusion.

2. The step strategy. The participants follow up step by step the consequences of a single possibility. If a problem includes a premise that makes a categorical assertion, then they may follow up its consequences. Otherwise, they make a supposition, i.e., an assumption, to start the strategy rolling. Here's an example. The premises were as follows in abbreviation:

A pink if and only if a black.

Either a black or else a gray.

A gray if and only if a blue.

Does it follow that: If not a pink then a blue?

The participant began by reading each premise aloud, and then said: "Assuming we have no pink", which is a supposition corresponding to the "if" clause of the conclusion. The participant repeated: "There is no pink", and crossed out the word "pink" in the first premise. The participant inferred: "So there is no black", thereby drawing the first of a series of conclusions concerning a single possibility. The participant crossed out the word, "black" in the first and second premises, and drew another conclusion: "There is gray," circling the word "gray" in the third premise. The participant drew another conclusion: "There is blue". Finally, the participant said, "Yes," to accept the conclusion, adding, "not pink and blue", and re-iterated the imprimatur, "yes".

The participants did not always use suppositions correctly. Given, say, a conclusion to evaluate, such as:

If not a red then a black

they sometimes made the supposition:

Suppose there's a black.

and were able to infer that there is not a red. They then responded that the conditional followed from the premises. They may have made the correct response, but they haven't truly shown than that the conditional follows from the premises. A conditional allows that the "then" clause can be true even when the "if" clause is false, and so the right way to proceed is to make a supposition of the "if" clause and to show that it leads to the truth of the "then" clause.

One variant of the step strategy was sophisticated. A few participants made a supposition of a *counterexample* to a conclusion, and then used the step strategy to pursue its consequences. For instance, given the problem:

Either a red or else a blue.

Either a blue or else a gray.

A gray if and only if a white.

Does it follow that: A red if and only if a white?

one participant reasoned as follows:

Assuming red and not white. [a counterexample to the conclusion]

Then not gray. [from the supposition and the third premise]

Then not red. [from the previous step and the second and first premises]

No, it is impossible to get from red to not red.

The main diagnostic signs of the step strategy are that reasoners start by stating a supposition or a categorical premise, and then infer a series of simple categorical conclusions, each concerning a single possibility. In cases where an inference yields more than one model, the conclusion often has a modal qualification, e.g., "possibly, there isn't a black marble", and any subsequent conclusions are themselves modal in the same way.

3. The compound strategy. Reasoners drew a compound conclusion from a pair of compound premises, where "compound" means that a sentence contains a connective. They sometimes made the inferences from diagrams of the premises. They expressed the conclusion either verbally or by drawing a new diagram, or both. By combining the conclusion with a new compound, they drew another compound conclusion, and so on, until they finally reached the answer to the question. Here's an example based on the premises:

A white if and only if a blue.

If a blue then a pink.

Either a pink or else a brown.

Does it follow that: a white or a brown?

The participant read aloud all the premises, and then drew a diagram to represent the first premise:

blue → white

The participant read the second premise again, wrote it down, and drew a conclusion (which was invalid) from these two premises:

If there's a pink then there's a white

The participant drew a diagram to represent this conclusion:

pink → white

From this conclusion and the third premise, the participant inferred:

If there's a brown then there isn't a white

Finally, from this conclusion, the participant inferred the answer to the question:

Either there's a brown or else a white.

In the compound strategy, one premise is used to construct models, and the other premise is used to update them. The combination of two compound premises can put a heavy load on working memory, especially when both premises have multiple models, and so it is not surprising that individuals sometimes draw a modal conclusion about only one possibility.

4. The chain strategy. This strategy was totally unexpected, and no mention of it appears to be in either the psychological or logical literature. The reasoner constructs a chain of conditionals leading from one clause in a conditional conclusion to the other

clause. This "chain" strategy resembles the step strategy, but has two crucial differences. First, reasoners do not announce that they are making a supposition. Indeed, they are not making a supposition, because they do not draw any intermediate conclusions. Second, they convert any premise that is not a conditional into a conditional, either verbally or in a diagram. These conversions include cases where a biconditional, such as:

There's a gray if and only if there's a red

is transformed into a conditional, such as:

If there isn't gray then there isn't a red.

The reasoner's aim is to ensure that the "then" clause in one conditional matches the "if" clause of the next conditional. A participant using this strategy began by drawing a separate diagram for each premise in the problem. The premises were depicted as I have shown in this statement of the problem:

A gray if and only if a red. $r \rightarrow g$

Either a red or else a white. $r \ X \ w$

A white if and only if a blue. $b \rightarrow w$

Does it follow that: If not a gray then a blue?

The participant, pointing at each diagram in turn, then said:

If not gray then not red.

If not red then white.

White comes from blue.

Yes.

The final "yes" was to acknowledge that the conclusion followed from the premises. This "chain" strategy is correct provided that reasoners construct a chain leading from the

"if" clause of the conclusion to its "then" clause. However, reasoners often worked incorrectly in the opposite direction. It is easier to make inferences from conditionals than from disjunctions, because conditionals have only one explicit mental model whereas disjunctions have at least two explicit mental models. Hence, the model theory predicts that chains of conditionals are much more likely than chains of disjunctions. Indeed, my colleagues and I have never observed anyone who developed a strategy in which the premises are converted into a chain of disjunctions.

5. The concatenation strategy. This strategy occurred only occasionally. The participants concatenated two or more premises in order to form an intermediate conclusion. They usually went on to use some other strategy, but sometimes they formulated their own conclusion by concatenating all the premises, and this conclusion was then used as the premise for an immediate inference yielding the required conclusion. For example, one participant argued from the premises:

A white and a blue.

A blue if and only if a black.

A black if and only if a red.

to the concatenation:

A white and a blue if and only if a black

and thence to the further concatenation:

A white and a blue if and only if a black if and only if a red.

At this point, the participant made an immediate inference to the required conclusion:

A white and a red.

The strategy accordingly depends on concatenating at least two premises into a single conclusion, and then either drawing such a conclusion, or else evaluating a given conclusion, if necessary by an immediate inference. The telltale sign of the strategy is that the participants join together premises and their connectives to form a single conclusion.

You might suppose the strategy depends on a formal procedure. On the contrary, it depends critically on mental models of possibilities. To see why, consider premises of the form:

A if and only if B.

Either B or else C.

C if and only if D.

where A, B, C, D, refer to the presence of different colored marbles in the box. Five different concatenated conclusions are possible, depending on the placing of the parentheses, e.g.:

A if and only if (B or else (C if and only if D))

((A if and only if B) or else C) if and only if D.

You might wonder which of them follows validly from the premises. In fact, none of them does. Four participants in one of our experiments used the concatenation strategy with these premises (see Van der Henst et al., 2002), and each of them spontaneously constructed this conclusion:

(A if and only if B) or else (C if and only if D).

It is the only concatenation out of the five possibilities that has the same mental models as those of the premises:

A B

C D

But, the conclusion is invalid, because its fully explicit models do not correspond to the mental models of the premises. Ten participants out of the twenty in this experiment used the concatenation strategy. On 82% of occasions, the resulting conclusions were compatible with the mental models of the premises, and nine of the ten participants concatenated more conclusions of this sort than not. Van der Henst and his colleagues accordingly concluded that concatenation depends on mental models.

Depending on the particulars of the experimental procedure, variation occurs in the frequencies with which the participants develop the different strategies. For example, when they have to formulate their own conclusions, they tend not to use the chain strategy. Conversely, the concatenation strategy is more frequent when they do have to formulate their own conclusions. The most frequent strategy in both cases, however, was the integrated diagram strategy. Participants mix strategies, and switch from one to another. Sometimes a switch occurs in the middle of a problem; sometimes from one problem to the next. There are no fixed sequences of steps that anyone invariably followed. Likewise, although the problems are all within the scope of sentential reasoning, the participants quite often went beyond its scope to draw intermediate conclusions about possibilities. Regardless of strategy, as a further experiment showed, problems that yield only one mental model are easier than those that yield two mental models, which in turn are easier than those that yield three mental models (Van der Henst et al., 2002).

The variety of strategies is not unique to reasoning on the basis of sentential connectives. It occurs when individuals reason about the relations between relations (Goodwin and Johnson-Laird, 2005a, b), and when they reason with quantifiers such as "all" and "some" (Bucciarelli and Johnson-Laird, 1999). It also occurs when they have to refute invalid conclusions based on sentential connectives. Given a conclusion that is consistent with the premises but that does not follow from them of necessity, their most frequent strategy is to try to construct a counterexample (Johnson-Laird and Hasson, 2003). With a conclusion that is not even consistent with the premises, the same studies showed that individuals tend to detect the inconsistency, that is, they grasp that the conclusion is impossible given the premises.

Conclusions

This article has outlined the model theory's account of deductive reasoning. Its main principles are:

1. Reasoning is based on models of possibilities, and each mental model represents what is common to a possibility.

2. As far as possible, a mental model is iconic: its structure represents the structure of the possibility that its represents.

3. Human reasoners tend to think about possibilities one model at a time.

4. Reasoning can proceed by conjoining the possibilities compatible with the different premises, but conjunctions of inconsistent models yield the null model.

5. Mental models are based on the principle of truth: they represent clauses in the premises only when they are true in possibilities. If individuals retain mental footnotes

about what is false then they can flesh out mental models into fully explicit models representing both what is true and what is false.

6. Semantic and pragmatic modulation affect the interpretation of connectives so they cannot be treated as strictly truth-functional.

7. Individuals develop different strategies for reasoning, e.g., they may use integrated diagrams to represent multiple possibilities, make steps from a single possibility, or think of a possibility that serves as a counterexample to a putative conclusion.

The theory has been applied to most domains of reasoning. They include reasoning with temporal relations (e.g., Schaeken, Johnson-Laird, and d'Ydewalle, 1996), spatial relations (e.g., Vandierendonck, Dierckx, and De Vooght, 2004), and other relations (e.g., Carreiras and Santamaría, 1997). It has also been applied to counterfactual reasoning (Byrne, 2005), causal reasoning (Goldvarg and Johnson-Laird, 2001), deontic reasoning (Bucciarelli and Johnson-Laird, 2005), reasoning from suppositions (Byrne and Handley, 1997), and reasoning about probabilities (e.g., Johnson-Laird, Legrenzi, Girotto, Legrenzi, and Caverni, 1999). Recently, it has begun to be extended to inductive reasoning, especially the sorts that occur in problem solving (Lee and Johnson-Laird, 2005a), in the revision of beliefs in the face of inconsistency (Johnson-Laird, Legrenzi, Girotto, and Legrenzi, 2000), in diagnoses (Goodwin and Johnson-Laird, 2005c), and in the reverse engineering of simple systems (Lee and Johnson-Laird, 2005b).

The chapter began with the effects of content on the selection task. Their discovery motivated the development of the model theory, and so you may be curious about what the theory has to say about the selection task. It postulates that individuals

rely on mental models of abstract claims, such as, "If a card has an 'A' on one side then it has a '2' on the other side." They think of the salient possibility represented in the one explicit mental model:

A 2

and they choose the 'A' card, and sometimes the innocuous '2' card too. To make the correct selections, they need to overrule the principle of truth in order to envisage a *counterexample* to the conditional:

A not 2

and then to choose the corresponding cards: A and 3. Most people fail. Any manipulation that makes counterexamples more salient, including the use of sensible contents or claims about what is permissible, yields an improvement in performance (see Johnson-Laird, 2001).

At the heart of the model theory is the assumption that individuals who have had no training in logic are able to make deductions. The theory does not abandon logic entirely (pace, e.g., Toulmin, 1958). Mental models relate to the model theory of logic, and to the logical principle that an inference is valid if there are no counterexamples to its conclusion. Individuals accordingly reason by constructing mental models of possibilities. These models are parsimonious in that they represent what is true, not what is false. The advantage is that individuals are able to make inferences that depend on more than one possibility. The disadvantage is that mental models can mislead individuals into thinking that they have grasped possibilities that in fact are beyond them.

Acknowledgements

The theory of mental models has developed as a result of the work of many researchers, see the Webpage maintained by Ruth Byrne and her colleagues:

www.tcd.ie/Psychology/Ruth_Byrne/mental_models/

I am grateful to these researchers for their help over the years. I am also grateful to the editors of this volume for their invitation to contribute this chapter, and for their helpful comments on an initial draft. The research was made possible in part by a grant from the National Science Foundation (Number 0076287) to study strategies in reasoning.

References

Baddeley, A. D. (1986) *Working Memory*. Oxford: Clarendon Press.

Barres, P., and Johnson-Laird, P.N. (2003) On imagining what is true (and what is false). *Thinking & Reasoning*, *9*, 1-42.

Barrouillet, P., Grosset, N., and Leças, J. F. (2000) Conditional reasoning by mental models: chronometric and developmental evidence. *Cognition, 75*, 237-266.

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

Bledsoe, W.W. (1977) Non-resolution theorem proving. *Artificial Intelligence*, 9. 1-35.

Boole, G. (1854) *An Investigation of the Laws of Thought on Which are Founded the Mathematical Theories of Logic and Probabilities*. London: Macmillan.

Braine, M.D.S. (1978). On the relation between the natural logic of reasoning and standard logic. *Psychological Review*, 85, 1-21.

Byrne, R.M.J. (2005) *The Rational Imagination: How People Create Alternatives to Reality*. Cambridge, MA: MIT.

Byrne, R.M.J., and Handley, S.J. (1997) Reasoning strategies for suppositional deductions. *Cognition*,62, 1-49.

Bucciarelli, M., and Johnson-Laird, P.N. (1999) Strategies in syllogistic reasoning. *Cognitive Science*, 23, 247-303.

Bucciarelli, M., and Johnson-Laird, P.N. (2005) Naïve deontics: a theory of meaning, representation, and reasoning. *Cognitive Psychology*, 50, 159-193.

Carreiras, M. and Santamaría, C. (1997) Reasoning about relations: Spatial and nonspatial problems. *Thinking & Reasoning*, 3, 191-208.

Cheng, P., and Holyoak, K.J. (1985) Pragmatic reasoning schemas. *Cognitive Psychology*, 17, 391-416.

Cosmides, L. (1989) The logic of social exchange: Has natural selection shaped how humans reason? *Cognition*, 31, 187-276.

Evans, J.St.B.T., Over, D.E., and Handley, S.J. (2005) Suppositions, extensionality, and conditionals: A critique of the mental model theory of Johnson-Laird & Byrne (2002). *Psychological Review*, in press.

García-Madruga, J.A., Moreno, S., Carriedo, N., Gutiérrez, F., and Johnson-Laird, P.N. (2001) Are conjunctive inferences easier than disjunctive inferences? A comparison of rules and models. *Quarterly Journal of Experimental Psychology*, 54A, 613-632.

Girotto, V., Mazzocco, A., and Tasso. A. (1997) The effect of premise order in conditional reasoning: a test of the mental model theory. *Cognition*, 63, 1-28.

Goldvarg, Y., and Johnson-Laird, P.N. (2001) Naïve causality: a mental model theory of causal meaning and reasoning. *Cognitive Science*, 25, 565-610.

Goodwin, G., and Johnson-Laird, P.N. (2005a) Reasoning about relations. *Psychological Review*, 112, 468-493.

Goodwin, G., and Johnson-Laird, P.N. (2005b) Reasoning about the relations between relations. *Quarterly Journal of Experimental Psychology*, 59, 1-23.

Goodwin, G., and Johnson-Laird, P.N. (2005c) Diagnosis of ambiguous faults in simple networks. *Proceeding of the Twenty-Seventh Annual Conference of the Cognitive Science Society*, Stresa, Italy. Mahwah, NJ: Erlbaum. In press.

Inhelder, B., and Piaget, J. (1958) *The Growth of Logical Thinking from Childhood to Adolescence*. London: Routledge & Kegan Paul.

Jeffrey, R. (1981) *Formal Logic: Its Scope and Limits*. Second edition. New York: McGraw-Hill.

Johnson-Laird, P.N. (1975) Models of deduction. In Falmagne, R.J. (Ed.) *Reasoning: Representation and Process in Children and Adults*. Hillsdale, NJ: Erlbaum. Pp. 7-54.

Johnson-Laird, P.N. (2001) Mental models and deduction. *Trends in Cognitive Science*, 5, 434-442.

Johnson-Laird, P.N. (2004) The history of mental models. In Manktelow, K., and Chung, M.C. (Eds.) *Psychology of Reasoning: Theoretical and Historical Perspectives*. New York: Psychology Press. Pp. 179-212.

Johnson-Laird (2005) Mental models, sentential reasoning, and illusory inferences. In Held, C., Knauff, M. and Vosgerau, G. (Eds.) *Mental Models: A Conception in the Intersection of Cognitive Psychology, Neuroscience and Philosophy*. Berlin: Elsevier. In press.

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

Johnson-Laird, P.N., and Byrne, R.M.J. (2002) Conditionals: A theory of meaning, pragmatics, and inference. *Psychological Review*, 109, 646-678.

Johnson-Laird, P.N., and Hasson, U. (2003) Counterexamples in sentential reasoning. *Memory & Cognition*, 31, 1105-1113.

Johnson-Laird, P.N., Legrenzi, P., Girotto, P., and Legrenzi, M.S. (2000) Illusions in reasoning about consistency. *Science*, 288, 531-532.

Johnson-Laird, P.N., Legrenzi, P., Girotto, V., Legrenzi, M., and Caverni, J-P. (1999) Naive probability: a mental model theory of extensional reasoning. *Psychological Review*, 106, 62-88.

Johnson-Laird, Legrenzi, P., and Sonino Legrenzi, M. (1972) Reasoning and a sense of reality. *British Journal of Psychology*, 63, 395-400.

Johnson-Laird, P.N., and Savary, F. (1999) Illusory inferences: A novel class of erroneous deductions. *Cognition*, 71, 191-229.

Lee, N.G.L., and Johnson-Laird, P.N. (2005a) Strategies in problem solving. Under submission.

Lee, N.G.L., and Johnson-Laird, P.N. (2005b) Synthetic reasoning and the reverse engineering of Boolean circuits. *Proceeding of the Twenty-Seventh Annual Conference of the Cognitive Science Society*, Stresa, Italy. Mahwah, NJ: Erlbaum. In press.

Manktelow, K.I. and Over, D.E. (1995) Deontic reasoning. In Newstead, S.E., and Evans, J.St.B.T. (Eds.) *Perspectives on Thinking and Reasoning: Essays in Honour of Peter Wason*. Mahwah, NJ: Erlbaum. Pp. 91-114.

Mill, J.S. (1874) *A System of Logic, Ratiocinative and Inductive: Being a Connected View of the Principles of Evidence and the Methods of Scientific Evidence*. Eighth Edition. New York: Harper. (First edition published 1843.)

Oaksford, M. and Chater, N. (1998) A revised rational analysis of the selection task: Exceptions and sequential sampling. In Oaksford, M. and Chater, N. (Eds.) *Rational Models of Cognition.* Oxford: Oxford University Press.

Ormerod, T. C., and Richardson, J. (2003) On the generation and evaluation of inferences from single premises. *Memory & Cognition*, 31, 467-478.

Osherson, D.N. (1974-6) *Logical Abilities in Children*, Vols. 1-4. Hillsdale, NJ: Erlbaum.

Peirce, C.S. (1931-1958) *Collected Papers of Charles Sanders Peirce*. 8 vols. Ed by Hartshorne, C., Weiss, P., and Burks, A. Cambridge, MA: Harvard University Press.

Popper, K. (1959) *The Logic of Scientific Discovery*. London: Hutchinson.

Rips, L.J. (1983). Cognitive processes in propositional reasoning. *Psychological Review*, 90, 38-71.

Rips, L. (1994) *The Psychology of Proof*. Cambridge, MA: MIT Press.

Schaeken, W.S., Johnson-Laird, P.N., d'Ydewalle, G. (1996) Mental models and temporal reasoning. *Cognition*, 60, 205-234.

Scriven, M. (1976) *Reasoning*. New York: McGraw-Hill.

Stenning, K., and Yule, P. (1997) Image and language in human reasoning: A syllogistic illustration. *Cognitive Psychology*, 34, 109-159.

Toulmin, S.E. (1958) *The Uses of Argument*. Cambridge: Cambridge University Press.

Van der Henst, J-B., Yang, Y., and Johnson-Laird, P.N. (2002) Strategies in sentential

reasoning. *Cognitive Science*, 26, 425-468.

Vandierendonck, A., Dierckx, V., and De Vooght, G. (2004) Mental model construction in linear reasoning: Evidence for the construction of initial annotated models. *Quarterly Journal of Experimental Psychology*, 57A, 1369-1391.

Walsh, C., and Johnson-Laird, P.N. (2004) Co-reference and reasoning. *Memory & Cognition*, 32, 96-106.

Wason, P.C. (1966) Reasoning. In Foss, B.M. (Ed.) *New Horizons in Psychology*. Harmondsworth, Middx: Penguin.

Wason, P.C., and Johnson-Laird, P.N. (1972) *The Psychology of Deduction: Structure and Content*. Cambridge, MA: Harvard University Press. London: Batsford.

Wason, P.C., and Shapiro, D. (1971). Natural and contrived experience in a reasoning problem. *Quarterly Journal of Experimental Psychology, 23,* 63-71.

Table 1: The procedures for conjoining sets of models illustrated with an example

The premises:

If Pat is here then Viv is here.

Mo is here or else Pat is here, but not both.

The first premise has the mental models:

Pat Viv

. . .

The second premise has the mental models:

Mo

Pat

The procedures:

Procedure 1 (which applies only to mental models): The conjunction of two mental models, such as: *A B*, and *B*, depends on the set of models from which the second model of *B* alone is drawn. If *A* occurs in one of these models, then its absence in the current model is treated as its negation. The conjunction is in effect a contradiction: *A B*, and *not-A B*. If *A* does not occur in the set of models from which *B* is drawn, then its absence does not yield a contradiction. Hence, the four conjunctions from the models yield the following equivalences:

Pat Viv	*and* Mo	is equivalent to:	Pat Viv	*and* Mo not-Pat
. . .	*and* Pat	is equivalent to:	not-Pat . . .	*and* Pat
Pat Viv	*and* Pat	is equivalent to:	Pat Viv	*and* Pat Viv
. . .	*and* Mo	is equivalent to:	Mo . . .	*and* Mo

Procedure 2. The conjunction of a pair of models containing respectively a proposition and its negation yield the *null* model, which represents contradictions:

Pat Viv *and* Mo not-Pat *yield* nil.

not-Pat . . . *and* Pat *yield* nil.

Procedure 3. The conjunction of a pair of models that do not contain a contradiction yields a model containing all the elements of both models (except that explicit content replaces implicit content):

Pat Viv *and* Pat Viv *yield* Pat Viv

Mo . . . *and* Mo *yield* Mo

Procedure 4. The conjunction of a pair of models in which at least one of them is the null model yields the null model, e.g.:

Pat Viv *and* nil *yield* nil.

The results:

The conjunction of the two sets of mental models above yields:

Pat Viv

Mo

These models support the valid conclusion:

Pat and Viv are here, or else Mo is here.

_