Mental Models in Discourse Processing and Reasoning
G. Rickheit and C. Habel (Editors)

13

MENTAL MODELS IN DEDUCTIVE, MODAL, AND PROBABILISTIC REASONING

Patrizia Tabossi, Dipartimento di Psicologia, Università di Trieste, Italy
Victoria A. Bell and P.N. Johnson-Laird, Department of Psychology, Princeton University, USA

ABSTRACT

This paper outlines the mental model theory of reasoning and compares it to theories of reasoning based on formal rules of inference. It shows how the model theory accounts for reasoning that leads to deductively valid conclusions, and it reports the results of an experiment that corroborates this account. This first experiment showed that inferences from premises that yield only one model are easier – they take less time and elicit fewer errors – than inferences from premises that yield multiple models. The model theory extends naturally to modal reasoning: A conclusion is *possible* if it is supported by one model of the premises, whereas a conclusion is necessary if it is supported by all the models of the premises. Hence, the theory predicts that it should be easier to infer what is possible than what is necessary. The opposite relation, however, should hold for the denial of conclusions about what is possible and what is necessary. The second experiment corroborated these predictions. A key assumption of the model theory is that individuals normally represent only what is true. This assumption led us – by way of a computer program implementing the theory – to a surprising phenomenon: The existence of illusory inferences with conclusions that seem obvious but that are egregious errors. The third experiment confirmed their existence in inferences about relative probabilities. The results also showed that the cause of the illusions is, in part, the failure to represent false contingencies. The experiment also confirmed that the probability of an event is inferred from the proportion of models in which it occurs.

INTRODUCTION

Reasoning is under intensive investigation by psychologists, artificial intelligencers, and other cognitive scientists. Some theorists have proposed that it depends on a memory for previous cases, or on general knowledge represented in the form of conditional rules or connectionist networks. Such theories, however, fail to explain the fact that human reasoners can make deductions that do not depend on general knowledge:

If it is a ubiflor then it is farmindarceous.
It is a ubiflor.
∴ It is farmindarceous.

This inference depends on a knowledge of language and especially a knowledge of 'if'. The central theoretical controversy is accordingly between two schools of thought about this sort of reasoning. The first school proposes that reasoning is a syntactic process depending on *formal rules of inference* akin to those of a logical calculus. The second school proposes that reasoning is a semantic process depending on *mental models* akin to the models that logicians invoke in formulating the semantics of their calculi. The controversy has been fruitful – it has led to improvements in experimental methodology and in the theories themselves. But it has been going on for a long time, and the time has come to try to settle it. Our aim in the present paper is to present some evidence that at least signals a beginning to the end of the controversy.

The plan of the paper is as follows: The rest of the introduction outlines the two sorts of competing theories – the formal rule and the model theories. The first part of this paper considers deductive reasoning, and reports an experiment that concerns reasoning with quantifiers, such as 'all' and 'some' and that examines the participants' reaction times to evaluate inferences. The second part considers modal reasoning – reasoning about conclusions that are possible or necessary – and reports an experiment that examines a key interaction predicted by the model theory. The experiment again uses both accuracy and reaction times as the main measures of performance. Part 3 considers probabilistic reasoning. It describes the phenomenon of illusory inferences – inferences that seem to support an 'obvious' conclusion that, in fact, is profoundly erroneous. It reports an experiment that demonstrates such illusions, but that also offers an insight into how to ameliorate them. Finally, the paper draws some conclusions about the controversy.

Formal rule theories

Some inferences, such as the example about ubiflors, can be made so rapidly and automatically that many theorists have assumed that the mind must be guided by formal rules in making them.

In the mid-1970s, several investigators for the first time formulated theories of human logical competence based on this assumption (see e.g. Osherson, 1974-6; Braine, 1978). The different theories postulated slightly different formal rules, and slightly different procedures for searching for derivations, but they had in common the notion that reasoning was similar to the process of constructing a formal proof. Philosophers, linguists, and artificial intelligencers have also defended this formal point of view (see e.g. Robinson, 1979; Macnamara, 1986; Sperber and Wilson; 1986; Pollock, 1989). The two major proponents of formal rules in psychology are the late Martin Braine (see e.g. Braine and O'Brien, 1991) and Lance Rips. Rips's (1994) PSYCOP theory is the first formal rule theory in psychology to cope with sentential connectives, such as 'if', 'or', and 'and', and with quantifiers, such as 'all' and 'some'. It is also the first formal rule theory to have been implemented in a computer program. Hence, we will treat PSYCOP as the paradigm case of a formal rule theory.

The first problem in developing a formal rule theory is to formulate psychologically plausible rules of inference. Rips, like most formal rule theorists, adopts the 'natural deduction' method from formal logic, and a key feature of this method is that it uses rules both to introduce connectives and to eliminate them. Table 1 presents a set of rules of the sort that Rips and other formal theorists adopt. The rule for eliminating 'if' shown in the table is often known as the rule for *modus ponens*, and it is this rule that Rips suggests underlies the ease of such deductions as:

If it is a ubiflor then it is farmindarceous.
It is a ubiflor.
∴ It is farmindarceous.

Another key feature of the 'natural deduction' method is the use of suppositions. These are assumptions that are made for the sake of argument, and that must be 'discharged' sooner or later if a derivation is to yield a conclusion. One way to discharge a supposition is to incorporate it in a conditional conclusion, which depends on the first rule for 'if' (the rule of *conditional proof*) in Table 1. Another way to discharge a supposition is to show that it leads to a contradiction and must therefore be false (according to the rule of *reductio ad absurdum*, which is not shown in Table 1). Thus, consider the following proof of an argument in the form known as *modus tollens*:

1. If it is a ubiflor then it is farmindarceous.
2. It is not farmindarceous.
3. It is a ubiflor. (A supposition)

The rule for *modus ponens* can be applied to premise 1 and to the supposition in order to derive:

4. It is farmindarceous. (*Modus ponens* applied to 1 & 3)

At this point, there is a contradiction between a sentence in the domain of the premises (It is not farmindarceous) and a sentence in the subdomain of the supposition (It is farmindarceous). The rule of *reductio ad absurdum* uses such a contradiction to negate – and thereby discharge – the supposition that led to the contradiction:

5. It is *not* a ubiflor.

Rips could have adopted a single rule for *modus tollens,* but the inference is harder for logically-untrained individuals than *modus ponens*, and so he assumes that it depends on the chain of inferential steps illustrated here.

Table 1: Some formal rules of inference for introducing and eliminating the sentential operator 'not', and the sentential connectives 'and', 'or', and 'if'. The expression 'A |- B' means that B can be derived in a proof from A.

Rules for introducing connectives	**Rules for eliminating connectives**
A ∴ not (not A)	not (not A) ∴ A
A B ∴ A and B	A and B ∴ A
A ∴ A or B	A or B not-B ∴ A
A |- B ∴ If A then B	If A then B A ∴ B

The second problem in developing a formal rule theory is to ensure that it is computationally viable. For example, unless the rule for introducing 'and' (see Table 1) is curbed, it can lead to such futile derivations as:

It is a ubiflor.
It is a farmindarceous.
∴ It is a ubiflor and it is a farmindarceous.

∴ It is a ubiflor and (it is a ubiflor and it is a farmindarceous).
∴ It is a ubiflor and (it is a ubiflor and (it is a ubiflor and it is a farmindarceous)).

and so on *ad infinitum*. Two sorts of rules are potentially dangerous: Those that introduce a connective and thereby increase the length of expressions; and those that introduce suppositions. A lesson from artificial intelligence is that programs can use a rule in two ways: Either to derive a step in a *forward* chain of inference from some assertions to a conclusion, or to derive a step in a *backward* chain from a conclusion to the subgoal of proving its required premises. The problem of curbing rules can be solved by using the potentially dangerous rules only in backward chains. Rips adopts this idea. PSYCOP accordingly has three sorts of rules: Those that it uses forwards, those that it uses backwards, and those that it uses in either direction.

PSYCOP can generate its own conclusions by using forward rules to derive them from the premises. In principle, it can use backwards rules if it guesses a putative conclusion. However, these rules are geared to the evaluation of *given* conclusions. The strategy that it then follows is to apply all its forward rules (breadth first) until they yield no new conclusions. It then checks whether the given conclusion is among the sentences that it has derived. If not, it tries to work backwards from the given conclusion, pursuing a chain of inference (depth first) until it finds the sentences that satisfy the subgoals or until it has run out of rules to try. Either it succeeds in deriving the conclusion or else it returns to an earlier choice point in the chain and tries to satisfy an alternative subgoal. Finally, if it fails all the subgoals, it gives up and responds that there is no valid conclusion. Unfortunately, it has no way to distinguish between a problem that really has no valid conclusion and a problem that has a valid conclusion that it is unable to derive.

One other aspect of PSYCOP should be mentioned: Its treatment of quantifiers, such as 'all' and 'some'. PSYCOP transforms quantified assertions into a format in which the work of quantifiers is performed by names and variables. The resulting expressions are akin to those used by automated theorem-provers in artificial intelligence, and Rips introduces various rules for matching one expression to another in these quantifier-free expressions. PSYCOP makes no surprising predictions, and it has not led to the discovery of any striking phenomena. Like other formal rule theories, its successes are more modest: It makes sense of a respectable body of data (see Rips, 1994).

The mental model theory

When individuals understand discourse, perceive the world, or imagine a state of affairs, then according to the mental model theory they construct mental models of the relevant situations. Reasoning is thus semantic, not syntactic, because reasoners build models of the relevant

situations based on their understanding of quantifiers and connectives, and, where relevant, on their general knowledge. They formulate an informative conclusion that is true in these models, and they assess its strength by searching for other models of the premises in which it is false (Johnson-Laird and Byrne, 1991). A mental model is, by definition, a representation that corresponds to a set of situations, and that has a structure and content that captures what is common to these situations. For example, an assertion such as, 'The ball is in the same place as the triangle' calls for a model of the form:

| o Δ |

in which 'o' denotes a model of the ball, 'Δ' denotes a model of the triangle, and the two vertical lines demarcate a place. In this case, two objects are represented by two corresponding mental tokens, with properties that represent the properties of the objects, and with a relation between the two tokens that represents the relation between the two objects.

The fundamental *representational* assumption of the model theory is that, in order to minimize the load on working memory, people normally represent explicitly only those situations that are true (Johnson-Laird and Byrne, 1991). This principle applies at two levels: Individuals represent only true possibilities; and they represent only the true components of these true possibilities. For example, given an exclusive disjunction, such as:

There is a ball or there is a triangle, but not both reasoners construct two alternative models to represent the two true possibilities:

o

Δ

where each row denotes a model of a separate possibility. Each model represents solely what is true. Hence, the first model represents that it is true that there is a ball, but it does not represent explicitly that in this case it is false that there is a triangle. Similarly, the second model represents that it is true that there is a triangle, but it does not represent explicitly that it is false that there is a ball. Reasoners may make a 'mental footnote' to keep track of what is false, but these footnotes are likely to be forgotten soon. Originally, Johnson-Laird and Byrne (1991) used square brackets as a special notation to denote these mental footnotes, but we will forego this notation here. Fully explicit models of the exclusive disjunction would represent the false components in each model:

o ¬Δ

¬o Δ

where '¬' represents negation and thus in this case falsity. The theory gives an analogous account of the other main sentential connectives, and Table 2 summarizes their models. It also shows the fully explicit models for these connectives.

The model theory provides the first unified account of deductive reasoning, modal reasoning, and probabilistic reasoning. A conclusion is deductively valid if it holds in all the models of the

premises. A conclusion is probable if it holds in most models of the premises. And a conclusion is possible if it holds in at least one model of the premises. This paper will present some new evidence in support of the model theory's account of all three of these main domains.

Table 2. The mental models and the fully explicit models for five sentential connectives: '¬' symbolizes negation, and '. . .' a wholly implicit model.

Connectives	Mental models	Fully explicit models
A and B:	A B	A B
A or B, not both:	A B	A ¬B ¬A B
A or B, or both:	A B A B	A ¬B ¬A B A B
If A then B:	A B . . .	A B ¬A B ¬A ¬B
If and only A then B:	A B . . .	A B ¬A ¬B

DEDUCTIVE REASONING

Previous studies of deductive reasoning have corroborated some of the predictions of the model theory. In the case of deductions based on both spatial reasoning and temporal reasoning, studies have shown that one-model problems are easier than multiple-model problems (see e.g. Johnson-Laird and Byrne, 1991; Schaeken *et al.*, 1996). The choice of problems in these studies enabled us to pit the model theory's predictions against those of formal rule theories, i.e. the one-model problems called for formal derivations that were longer than those called for by the multiple-model problems.

Studies of deductions based on sentential connectives, such as 'not', 'if', 'and', and 'or', have also shown that one-model deductions are easier than multiple-model deductions (Johnson-Laird and Byrne, 1991). In a recent unpublished study, Fabien Savary and Johnson-Laird video-taped

individuals as they carried out sentential reasoning armed with pencil and paper. They often devised their own idiosyncratic diagrams. Given the premises:

Either there is a grey marble or else there is a brown marble, but not both.
There is a brown marble if and only if there is a white marble.
Either there is a white marble or else there is a blue marble, but not both.

one person, for example, drew a diagram in which each row represents a possible state of affairs:

blue grey
white brown

Another person drew a vertical line and then added the colors in the following arrangement:

| white
grey | brown
blue |

Such diagrams are isomorphic to mental models of the premises (cf. Table 2). Studies of reasoning with quantifiers have also examined the difference between one-model and multiple-model problems for syllogisms, such as:

Some of the parents are scientists.
All the scientists are drivers.
What follows?

These studies showed that one-model syllogisms were easier than multiple-model syllogisms (see e.g. Bara *et al.*, 1995). Other studies have observed the same effect for doubly-quantified premises (Johnson-Laird *et al.*, 1989), such as:

Some of the artists are in the same place as all of the beekeepers.
All of the beekeepers are in the same place as all of the chemists.
What follows?

An important feature of all of these studies of spatial, temporal, sentential, and quantified reasoning is that reasoners' erroneous conclusions tend overwhelmingly to correspond to conclusions that are supported by an initial model of the premises.

One sort of study of quantified reasoning has been conspicuously missing. Few studies have examined the latencies of the participants' responses; syllogisms and multiply-quantified problems elicit too many errors for this method to be used with much success. The aim of our first experiment was to fill this gap. We therefore used a new sort of problem that was simple enough for us to make meaningful measures of reaction times. These problems are based on two premises, but only one of the three terms in the premises is quantified, e.g.:

Carla is in the same place as Rosaria.
Rosaria is not in the same place as any of the friends.
∴ Carla is not in the same place as any of the friends.

The participants' task was to evaluate whether or not the conclusion follows from the premises.

Experiment 1

The experiment was carried out at the University of Bologna, and its materials are translated from the original Italian. It investigated three sorts of problems: One-model problems, multiple-model problems with a valid conclusion, and multiple-model problems with no valid conclusion. The previous problem is an example of a one-model problem. The premises yield the following model:

| Carla Rosaria | friend friend friend |

where the vertical lines demarcate places, and the number of friends is arbitrary, but small. This model supports the given conclusion, and there is no alternative model of the premises in which the conclusion is false. An example of a multiple-model problem with a valid conclusion is:

Carla is in the same place as Rosaria.
Rosaria is not in the same place as some of the friends.
∴ Carla is not in the same place as some of the friends.

The similarity between this problem and the previous one-model problem is greater in the original Italian, where 'any' is the singular 'alcuno' and 'some' is its plural 'alcuni'. These premises support the same initial model as before:

| Carla Rosaria | friend friend friend |

This model is consistent with the stronger conclusion:

Carla is not in the same place as any of the friends.

But, this conclusion is refuted by an alternative model of the premises:

| Carla Rosaria friend | friend friend friend |

The two models together support the given conclusion:

Carla is not in the same place as some of the friends.

and there is no model of the premises that refutes this conclusion. An example of a multiple-model problem with no valid conclusion is:

Carla is not in the same place as some of the friends.
Some of the friends are in the same place as Rosaria.
∴ No valid conclusion.

The premises elicit the initial model:

| Carla | friend friend friend Rosaria |

which supports the conclusion:

Carla is not in the same place as Rosaria.

An alternative model of the premises refutes this conclusion:

| Carla Rosaria | friend friend friend |

and indeed there is no valid conclusion interrelating Carla and Rosaria (apart from the empty tautology: Carla is, or is not, in the same place as Rosaria).

The model theory predicts that reasoners should find it easier to draw valid conclusions to the one-model problems than to the multiple-model problems: They should make more correct responses and they should respond faster. The most difficult items should be the multiple-model problems with no valid conclusions, because the correct response can be established only by constructing both models. In addition, the correct response is 'no', and negatives are a well-known cause of difficulty (see e.g. Wason, 1959; Clark, 1969).

Method

Design: The participants acted as their own controls and evaluated 16 one-model problems with a valid conclusion, 16 multiple-model problems with a valid conclusion, and 16 multiple-model problems with no valid conclusion. Each set of 16 problems was derived from underlying sets of four logically distinct problems, which are shown in schematic form in Table 3. As the Table shows, each schema has two premises based on the relation 'in the same place as', and contains three terms, a, b, c, one of which is quantified. We used each of the 12 underlying schemas to form two problems depending on the order of the two premises. One order yields problems with terms in the figure: a-b, b-c:

a is in the same place as b.
b is in the same place as some of the c.

The other order yields problems with terms in the figure: b-a, c-b:

b is in the same place as some of the a.
c is in the same place as b.

Each of the resulting 24 schemas was combined with two putative conclusions so that half the problems had a conclusion to which the correct response was 'yes', and half the problems had a conclusion to which the correct response was 'no'. In order to elicit the 'yes' responses, the one-model and the multiple-model problems with valid conclusions were presented with their valid conclusions, and the multiple-model problems with no valid conclusion were presented with 'No valid conclusion'. In order to elicit the 'no' responses, half of the one-model and the multiple-model problems with valid conclusions were presented with 'No valid conclusion', and half of them were presented with invalid conclusions, and the problems with no valid conclusion were presented with an invalid conclusion. Finally, we allocated the lexical contents to the 48 problems at random, and divided them into four sets of 12 problems in which the three sorts of problems, the figure of the problems, and the number of 'yes' and 'no' responses were balanced equally. Each participant was presented with the four sets of problems in a different random order.

Participants: 16 undergraduates at the University of Bologna (12 women and 4 men) took part voluntarily in the experiment, which lasted about 30 min. None of them had participated in an experiment of this sort before, nor had any of them taken courses in logic.

Table 3. The underlying structures of the three sorts of problems in Experiment 1 together with their mental models. 'S' denotes the relation 'is in the same place as', lower case letters denote individuals, and vertical lines in models demarcate places, e.g. 'a S some b' denotes a premise, such as 'Carla is in the same place as some of the friends'. Conclusions in parentheses are of the form used to elicit a 'no' response.

The three sorts of problems		
One model	**Multiple-model with valid conclusions**	**Multiple model with no valid conclusions**
a S b b S some c ∴ S some c (∴ a not S some c)	a not S b b S some c ∴ a not S some c (No valid conclusion)	a S some b Some b not S c No valid conclusion (∴ a S c)
| a b c c c |	| a | b c c c | | a c | b c c |	| a b b b | c | | a b b c | b |
a S b b not S any c ∴ a not S any c (∴ a S some c)	a S b b not S some c ∴ a not S some c (∴ a S some c)	a not S some b Some b S c No valid conclusion (∴ a not S c)
| a b | c c c |	| a b | c c c | | a b c | c c |	| a | b b b c | | a b b c | b |
Some a S b b S c ∴ Some a S c (No valid conclusion)	Some a S b b not S c ∴ Some a not S c (∴ None a S c)	a not S any b None b S c No valid conclusion (∴ a S c)
| a a a b c |	| a a a b | c | | a a b | a c |	| a | b b b | c | | a c | b b b |
None a S b b S c ∴ None a S c (No valid conclusion)	Some a not S b b S c ∴ Some a not S c (No valid conclusion)	None a S b b not S c No valid conclusion (∴ Some a S c)
| a a a | b c |	| a a a | b c | | a a | a b c |	| a a a | b | c | | a a a c | b |

Materials: The set of 48 problems (premises and conclusion) were each based on the spatial relation, 'in the same place as', and the terms in the problems were common Italian first names and frequent Italian common nouns.

Procedure: The participants in the experiment were tested individually. They sat in front of a computer screen in a quiet room. The experimenter explained that their task was to decide whether or not a conclusion followed from the premises in a series of problems, and that a conclusion followed from the premises if it had to be true given that the premises were true. They were told to read the premises, and then to press the 'yes' button as quickly as possible if the conclusion followed from the premises, and to press the 'no' button as quickly as possible if the conclusion did not follow from the premises. There were four practice trials before the start of the experiment proper; two required a 'yes' response and two required a 'no' response. After a 150 msec. warning tone, each problem appeared on the screen as the computer-controlled timer started. The problem was presented with the two premises separated from each other by a blank line, and the conclusion separated from the second premise by two blank lines. The participant responded by pressing a button. Half the participants pressed the 'yes' button with their dominant hand, and half the participants pressed the 'no' button with their dominant hand. Four participants were left-handed. The participant's response stopped the timer. There were 5 sec. intervals between trials. There was a short interval between the presentation of the four lists of problems while the experimenter prepared the next list. The computer recorded the participant's response, whether or not it was correct, and its latency.

Results and discussion: Table 4 presents the percentages of correct 'yes' and 'no' responses to the three sorts of problems, and the latencies for the correct responses. The nature of the problems had a marked effect on accuracy ($F(2,34) = 10.975$, $p < 0.0003$). The one-model problems elicited more correct responses than the multiple-model problems with or without valid conclusions ($F(1,34) = 10.721$, $p < 0.003$, a designed comparison), whereas there was no reliable difference between multiple-model problems with and without valid conclusions ($F(1,34) = 1.600$, n.s, a designed, though non-orthogonal, comparison). There were no other significant effects on accuracy.

The pattern of latencies corroborated the accuracy results, except that the 'yes' responses (mean 10.3 secs) were reliably faster than the 'no' responses (mean 11.1 secs; $F(1, 17) = 8.390$, $p < 0.01$). The nature of the problems yielded a robust effect ($F(2,34) = 19.982$, $p < 0.0002$). One-model problems were responded to faster than the multiple-model problems with or without valid conclusions ($F(1,34) = 31.212$, $p < 0.0001$, a designed comparison). There was no reliable difference between the multiple-model problems with and without valid conclusions ($F(1,34) = 0.053$, n.s., a designed, non-orthogonal, comparison).

Table 4. The percentages of correct 'yes' and 'no' responses to the three sorts of problems in Experiment 1 and in parentheses the latencies of the correct responses in seconds.

	The three sorts of problems		
	One model	**Multiple model with valid conclusions**	**Multiple model with no valid conclusions**
'Yes' responses	86 (9.1)	72 (11.1)	60 (10.8)
'No' responses	78 (9.7)	62 (11.6)	63 (12.0)
Overall	82 (9.4)	67 (11.3)	61 (11.4)

The results corroborated the main prediction of the model theory. The reasoners drew a greater number of correct conclusions from the one model problems than from either the multiple-model problems with valid conclusions or the multiple-model problems with no valid conclusions. In addition, they were also faster to make the correct response to the one- model problems than to either of the other sorts of problems. In other words, there was no trade off between accuracy and latency. We draw two morals from these results: First, with materials that elicit a preponderance of correct conclusions, it is possible to use the speed of reasoning as a dependent variable. Second, speed and accuracy strongly suggest that individuals build models of the premises in order to reason.

MODAL REASONING

There are no full-fledged formal rule theories of modal reasoning, that is, reasoning about what is possible and what is necessary. One way to extend such theories so that they deal with the domain is to introduce modal rules of inference (see Osherson, 1974-6, who formulates some rules for deriving certain modal conclusions from certain modal premises). Logicians have formulated various modal logics (see e.g. Chellas, 1980), but within them some simple inferences require derivations that are too complicated to be psychologically plausible. Consider, for example, the following transparent inference:

Ewing is in the game or Starks is in the game, or both.
∴ It is possible that both Ewing and Starks are in the game.

No formal rule should be of the form:

p or q
∴ possibly (p and q)

because, as Geoffrey Keene (personal communication) has pointed out, if p implies *not* q, then the result of the rule would be a conclusion stating that a self-contradiction is possible. The mere fact that an inclusive disjunction is true does not imply that both disjuncts can be simultaneously satisifed, e.g. the disjunction "2 + 2 = 4, or 2 + 2 = 5, or both" is true because the first of its disjuncts is true. The first step in a formal derivation is therefore to translate the premise into one having the following logical form:

(p or q) and not (p necessarily implies not q)

The required conclusion can now be derived using modal logic (see T18.31 of Lewis and Langford, 1932, p. 163). Hence, what is simple for people can be highly complicated in formal logic. In contrast to formal rule theories, the model theory extends naturally to deal with modal reasoning. A state of affairs is possible – it *can* happen – if it occurs in at least one model of the premises; and a state of affairs is necessary – it *must* happen – if it occurs in all the models of the premises. Conversely, a state of affairs is not necessary – it is *not* the case that it must happen – if it fails to occur in at least one model; and a state of affairs is not possible – it is not the case that it can happen – if it fails to occur in all the models of the premises (see Johnson-Laird, 1994). In other words, a single example establishes a claim about what is possible, and a single counterexample suffices to refute a claim about what is necessary; in contrast, all cases must be counterexamples to refute a claim about what is possible, and all cases must be examples to establish a claim about what is necessary.

The theory therefore makes a strong prediction about modal reasoning. It predicts an interaction between the modality of the conclusion ('possible' versus 'necessary') and its polarity ('affirmative' versus 'negative'). On the one hand, conclusions about what is possible should be easier to draw – faster and more accurate – than conclusions about what is necessary. On the other hand, conclusions about what is *not* possible should be harder to draw than conclusions about what is *not* necessary. This key interaction is the central prediction of the model theory about modal reasoning.

Formal rule theories are unlikely to predict the interaction. There is no role in them for examples or for counterexamples. They establish that a claim is not necessary by failing to find a derivation for it. Hence, such theories – if they are ever formulated – are likely to predict that refutations of conclusions about what is possible or about what is necessary should be harder than proving such conclusions.

Our first test of the key interaction failed, probably because the experiment required the participants to respond that something was 'possible' when, in fact, it was obviously necessary. We therefore designed Experiment 2 to test the interaction in a way that avoided this

pragmatically odd response. The participants read two premises about the players in a game of one-on-one basketball, i.e. games in which there are only two players, who play against each other. Thus, of the four players referred to in the following premises, two are in the game, and two are out of the game:

1. If Ewing is in then Starks is in.
   If Oakley is in then Johnson is out.
   Can Starks be in the game?

According to the model theory (see Table 2), the first premise elicits the models:

Ewing Starks
. . .

where 'Ewing' denotes a model of Ewing in the game, and 'Starks' denotes a model of Starks in the game. Reasoners should also make a mental footnote that the explicit model exhausts the models in which Ewing occurs, i.e. any model containing Ewing also contains Starks. The converse is not true, i.e. Starks can occur in models in which Ewing does not occur. To answer the question, reasoners need verify only that the explicit model is consistent with the second premise, i.e. it is a member of the set of models of the second premise. The second premise elicits the models:

Oakley ¬ Johnson
. . .

where '¬ Johnson' denotes a model of Johnson out of the game, i.e. not in the game. Reasoners who go no further will judge that the model containing Ewing and Starks is consistent with these models, because these two players can occur in one of the cases represented by the wholly implicit model (denoted by the ellipsis). In fact, they will be correct, because if the second set of models is fleshed out explicitly, they are:

Oakley ¬ Johnson
¬ Oakley Johnson
¬ Oakley ¬ Johnson

Granted that two players must be in the game, the last of these three models represents the case where both Ewing and Starks are playing:

¬ Oakley ¬ Johnson Ewing Starks

Now, let us consider the same premises but with the question concerning necessity:

Must Starks be in the game?

In this case, reasoners need to verify that all possible models of the premises contain Starks. Given that the first premise allows that Starks can play without Ewing, Starks can be added to each model of the second premise:

Oakley	¬ Johnson	Starks
¬ Oakley	Johnson	Starks
¬ Oakley	Johnson	Starks

and to make up the team of two players, Ewing must be added to the third of these models. In sum, the premises are consistent with three possible games:

Oakley		Starks	
	Johnson	Starks	
		Starks	Ewing

It follows that Starks must be in the game. If reasoners construct these models, then they can respond, 'Yes,' to the question for the right reasons. An alternative strategy is to try to construct a model in which Starks is out. Consider the second set of models:

Oakley	¬ Johnson
¬ Oakley	Johnson
¬ Oakley	¬ Johnson

In the first case, if Starks is out, then Ewing is the only player left to be in. But, if Ewing is in, then so is Starks; and the result would be an illegal game with three players instead of two. Hence, Starks must be in. The same argument applies *mutatis mutandis* to the second model. And, as we have seen, Starks and Ewing must complete the third model. Once again, reasoners have to consider all three models in answering the question using this strategy.

To create a problem to which the correct answers to the two modal questions are negative, one method is to construct the *dual* of the previous problem 1, i.e. to change 'in' to 'out', and *vice versa*. The resulting dual is:

2. If Ewing is out then Starks is out.
   If Oakley is out then Johnson is in.

This problem has the following three fully explicit models:

¬ Ewing	¬ Starks	Oakley	Johnson
Ewing	¬ Starks	¬ Oakley	Johnson
Ewing	¬ Starks	Oakley	¬ Johnson

It is therefore impossible for Starks to play, and so both the possible and necessary questions have negative answers. Given the necessary question:

Must Starks be in the game?

reasoners are likely to construct the most salient model of the first premise:

¬ Ewing ¬ Starks

and then to establish that the second premise allows both Oakley and Johnson to play. The answer to the question is accordingly, 'No'. In contrast, given the possible question:

Can Starks be in the game?

reasoners must now consider all three possible models of the premises in order to answer 'No' correctly.

Problems 1 and 2, which are based on conditional premises, can also be expressed using inclusive disjunctions, because in this domain an assertion of the form:

If p then q

is equivalent to one of the form:

not-p and/or q

The disjunctive equivalents of problems 1 and 2 are thus:

1'. Ewing is out and/or Starks is in.
Oakley is out and/or Johnson is out.

2'. Ewing is in and/or Starks is out.
Oakley is in and/or Johnson is in.

Different models are likely to be salient when the problems are expressed using disjunctions. However, the theory still predicts the key interaction. In summary, the affirmation of a possibility and the denial of a necessity are both established by a single model, whereas the denial of a possibility and the affirmation of a necessity are both established by three models. Hence, the former should be inferred faster and more accurately than the latter.

Experiment 2

The experiment investigated problems based on conditionals, e.g.:

If Ewing is in then Starks is in.
If Oakley is in then Johnson is out.

and equivalent problems based on inclusive disjunctions, e.g.

Ewing is out and/or Starks is in.
Oakley is out and/or Johnson is out.

We used 'and/or' to express inclusive disjunction, because a pilot study showed that participants were confused by the tag, 'or both', as in 'Ewing is out or Starks is in, or both', where it was sometimes taken to mean that both players were in the game. After the participants had read the premises, they had to answer a question about either a possibility:

Can Starks be in the game?

or a necessity:

Must Starks be in the game?

All the problems have three models and rule out three models as impossible. Each problem was presented twice (with four different names), once with a 'can' question about a particular player and once with a 'must' question about the corresponding player. For half the problems, this player was necessary – though this fact was not obvious, and so the correct answer was affirmative to both questions. For the other half of the problems, this player was impossible – though again this fact was not obvious, and so the correct answer was negative to both questions. However, when the participants answer the 'can' question, they should stop considering models as soon as they can respond 'yes'. Likewise, when they answer the 'must' question, they should stop considering models as soon as they can respond 'no'. Hence, the experiment should avoid the pragmatic

difficulty that plagued our preliminary study: The participants should be unlikely to realize that a possible player is necessary, and that an unnecessary player is impossible.

Method

Design: There were eight distinct problems based on whether the premises were conditionals or disjunctions, the question was about a possibility or a necessity, and whether the correct answer was affirmative or negative. Table 5 presents the four problems with a necessary player and the about here models that they elicit. The four problems that yield an impossible player are their duals obtained by changing 'in' to 'out', and vice versa. The participants served as their own controls and carried out four versions of each of the eight problems – a total of 32 problems. Each of these problems concerned four different individuals. The participants were randomly assigned to do the problems in either one random order or its opposite order.

Participants: Twenty Princeton University undergraduates were recruited through a pool based on an introductory psychology course. All the students received class credit for one hour of participation. None of them had any formal training in logic.

Materials: The eight logically distinct sorts of problems were used to construct 32 sorts of problems by assigning to them distinct sets of four players' names. We used common two-syllable first names, each containing five letters. The problems concerned players who were 'in' or 'out', and overall the conditional premises had the same number of 'ins' and 'outs' as the disjunctive premises. The only difference between the 'possible' conditions and the 'necessary' conditions was whether 'can' or 'must' occurred in the question (and the four names of the players).

Procedure: The problems were presented on an Apple laptop computer running the MacLab reaction time program, and the participants entered their responses with key presses. The 'Y' key was labeled 'YES', the 'N' key was labeled 'NO', and the 'H' key was labeled '?' to mean 'I don't know'. The 'H' key was also used to bring up the next screen. Each trial began with a screen with the words 'Press H to begin'. The two premises were presented simultaneously, one below the other, on the next screen. The participants were told to read the premises before pressing the key to get the question. The premises stayed on the screen while the question appeared underneath. After the participants made their response, the correct answer appeared below the question.

Before the experiment, the participants read the instructions, which explained the nature of the problems, the form of the premises, and the two sorts of question. There were two examples of problems. Finally, the instructions stated: 'Take time to think, as accuracy is more important than

speed.' The participants' responses to the problems and their latencies were recorded from the time of the key press to uncover the question until the key press to respond.

Table 5. The four problems in Experiment 2 with a necessary player (shown in bold). The table shows the mental models and the fully explicit models, where they differ from the mental models. The four problems with an impossible player are obtained by switching 'in' for 'out' and vice versa.

Premises	Mental models	Fully explicit models
If A in then B in. If C in then D out.	A **B** C ¬D . . .	A **B** ¬C ¬D ¬A **B** C ¬D ¬A **B** ¬C D
A out and/or B in. C out and/or D out.	¬A **B** C ¬D ¬A **B** ¬C D A **B** ¬C ¬D	Same as the mental models
If A out then B out. If C in then D out.	¬A ¬B C ¬D . . .	**A** ¬B C ¬D **A** B ¬C ¬D **A** ¬B ¬C D
A in and/or B out C out and/or D out	**A** ¬B C ¬D **A** ¬B ¬C D **A** B ¬C ¬D	Same as the mental models

Results and discussion: Table 6 presents the percentages of correct responses to the four sorts of problems (affirmative possibility, negative possibility, affirmative necessity, and negative necessity), and the latencies of the correct responses (in secs). There was no reliable difference in accuracy or speed between the conditional and disjunctive problems, and so we have pooled their results. The participants were more accurate, however, in responding 'yes' than in responding 'no' (Wilcoxon Test, $z = 1.993$, $p < 0.05$), the difference presumably reflected the well-established difference between affirmatives and negatives (see e.g. Wason, 1959; Clark, 1969). The difference in latency between the two sorts of questions was marginally significant, i.e. 'can' elicited faster responses than 'must', but there was no difference in accuracy. These differences are

much less important than the key interaction. It was corroborated by the pattern of correct responses: The participants made fewer errors on affirmative possibilities than on negative possibilities, but they made more errors on negative possibilities than on negative necessities. Of the 20 participants, 14 followed the prediction, one went against it, and there were five ties (Wilcoxon Test, $n = 15$, $z = 3.304$, $p < 0.001$). An analysis of the results by materials corroborated the interaction: The analysis yielded the highest significance possible for four items per condition (Wilcoxon Test, $z = 1.826$, $p < 0.04$).

Table 6. The percentages of correct responses to the four sorts of problems in Experiment 2 and in parentheses the latencies of the correct responses in secs.

	Possible questions	**Necessary questions**	**Overall**
'Yes' responses	91 (18.0)	71 (25.6)	81 (21.8)
'No' responses	65 (22.3)	81 (22.7)	73 (22.5)
Overall	78 (20.1)	76 (24.1)	77 (22.0)

The key interaction was also corroborated by the response times. The participants were faster to respond 'yes' correctly to questions about possible players than to questions about necessary players, but they were slower to respond 'no' correctly to questions about possible players than to questions about necessary players. Out of the 20 participants, 17 showed the predicted interaction in their response times data (Wilcoxon Test, $z = 2.912$, $p < 0.004$). The correct 'no' responses about a possible player were faster than we expected, but the pattern of errors suggests that there may be a speed accuracy trade-off for the problems in this condition.

In general, the results bear out the model theory's prediction of a key interaction: Reasoners are faster and more accurate in inferring that a player is possible as opposed to necessary, but they are faster and more accurate in inferring that a player is not necessary as opposed to not possible. This robust pattern is only to be expected if reasoners infer that a state of affairs is possible by finding an example of it among the models of the premises, but infer that state of affairs is necessary by finding it in all the models of the premises. Conversely, they infer that a state of affairs is not necessary by finding a counterexample to it among the models of the premises, but infer that a state of affairs is impossible by finding that it does not occur in any of the models of the premises. A theory based on formal rules of inference may be able to accommodate the interaction, but, as we suggested, the accommodation will not be easy. Neither examples nor counterexamples play any role in theories of deductive reasoning based on formal rules of inference (see e.g. Rips, 1994).

PROBABILISTIC REASONING

Formal rule theories have yet to be formulated as accounts of how people make probabilistic inferences. Once again, however, the model theory extends naturally to explain naive reasoning about probabilities, i.e. reasoning not explicitly based on the probability calculus. Numerical judgments of probability may sometimes be based on the availability of models (see Tversky and Kahneman, 1973). Other judgments may call for models to be linked to numerical representations of odds or probabilities. The model theory, however, can account for a variety of naive judgments. It postulates:

1. *The 'frequency' assumption: The more models of the premises in which an event occurs, the greater its probability should be judged to be.* In other words, events that occur in many models should be judged to be likely, whereas those that occur in only a few models should be judged to be unlikely. Consider, for example, the following problem, which we have investigated in a recent study (Johnson-Laird *et al.*, 1996):

> There is a box in which there is a green ball, or a red ball, or both.
> Given the preceding assertion, according to you, what is the probability of the following situation?
> In the box there is at least a green ball.

Readers may suppose that the only sensible answer is that it is impossible to give an estimate, because the probability could be anywhere between 0 and 1. Naive reasoners, however, are not reluctant to make specific estimates. The model theory postulates that the premise yields three models:

```
green
        red
green   red
```

These models will yield an estimate given that reasoners make a further assumption:

2. *The 'equiprobability' assumption: Each model represents a set of situations, and, in the absence of any evidence to the contrary, the sets are equiprobable* (cf. the analogous principle of 'indifference', see Hacking, 1975).

Equiprobability is a principle that reasoners should assume by default, i.e. unless they have evidence to the contrary. They are unlikely to assume that each candidate in a Presidential election

has an equal probability of winning. A knowledge that one candidate is, say, less popular than the others overrules the assumption of equiprobability. With the example about red and green balls, however, the assumption of equiprobability implies that each of the three models is equiprobable. Reasoners should accordingly judge that the probability of:

In the box there is at least a green ball.

is 66% because the green ball occurs in two out of the three models. Our results corroborated this and a variety of other predictions concerning disjunctions, conditionals, and conjunctions.

How do people judge which of two states of affairs is more likely? There are two possible ways. One way is to establish that one state of affairs occurs in a proper subset of all the models in which the other state of affairs occurs, e.g. a blue ball is more probable than a grey ball if a grey ball occurs only in a proper subset of the models in which a blue ball occurs. This principle yields correct judgments provided that reasoners consider all the possible models in which a blue ball or grey ball occurs. Consider, for example, an assertion of the form:

In the box, if there is a grey ball then there is a blue ball.

Reasoners who construct the models:

grey blue
...

will infer that the grey ball and the blue ball are equiprobable. Those who make an explicit interpretation of a 'one way' conditional:

grey blue
blue
...

will judge that a blue ball is more probable than a grey ball, because a grey ball occurs only in a proper subset of the models in which a blue ball occurs.

Together, the frequency assumption and the equiprobability assumption imply a less stringent way of making judgments of relative probabilities (see Johnson-Laird and Savary, 1996):

If event A occurs in more models than event B, then event A is more probable than event B.

This principle follows at once from the two assumptions: If the probabilities of states of affairs are judged from their frequency in models and if each model represents an equiprobable state of affairs, then one event is more probable than another when it occurs in more models. This method is risky, because the equiprobability assumption may be overruled by specific knowledge. Consider which is more likely, for instance, a green marble or a red marble, in the following set of models:

red
green
blue
green

Relative frequency yields the judgement that a green marble is more likely than a red marble, because it occurs in more models. If the equiprobability assumption is false, however, then the model in which a red marble occurs could be much more likely than all the other models put together. In cases of this sort, the more stringent subset principle would protect reasoners from error: They would respond that it is impossible to judge the relative probability, because neither marble occurs in a proper subset of the models in which the other marble occurs.

In previous studies, Johnson-Laird and Savary (1996) have corroborated the use of both the frequency and the equiprobability assumptions. They showed that reasoners make judgments of the relative probabilities of two events according to their relative frequency in models. Their study exploited a phenomenon of importance in its own right: The existence of illusory inferences. We will outline this phenomenon as it occurs in naive probabilistic reasoning.

Illusory inferences

The fundamental representational assumption of the model theory, which we described earlier, is that reasoners normally represent only what is true. Hence, given an exclusive disjunction, such as:

There is a king in the hand or else there is an ace in the hand

they tend to build the following two models (see Table 2):

king
ace

where each line denotes a separate model. The representational assumption has an unexpected consequence that we discovered by accident in the output of a computer program implementing the theory. Certain inferences have initial models that support a wholly erroneous conclusion. Consider the following example:

> Suppose that *only one* of the following assertions is true about a specific hand of cards:
> There is a king in the hand or there is an ace in the hand, or both.
> There is a queen in the hand or there is an ace in the hand, or both.
> Which is more likely to be in the hand: the king or the ace?

Most people infer that the ace is more likely (Johnson-Laird and Savary, 1996). This response is predicted by the model theory. The first disjunction yields the models:

```
king
          ace
king      ace
```

and the second disjunction yields the models:

```
queen
          ace
queen     ace
```

The main exclusive disjunction that combines the two premises calls for a list of the true possibilities (see Table 2). Hence, the models of the premises as a whole are:

```
king
          ace
king      ace
     queen
          ace
     queen ace
```

The ace occurs in more models than the king, and so, granted that participants base their estimates on the frequency assumption and the equiprobability assumption, they should judge that the ace is more probable than the king.

This response is totally wrong. It is based solely on what is true in the models. The instruction, 'Only one of the following assertions is true', implies that the other assertion is false. (Perhaps it

has no definite truth value – a possibility that we will consider in a moment.) Hence, when the first disjunction is true, then the second disjunction is false. The second disjunction (queen or ace) is false when there is neither a queen nor an ace. Conversely, when the second disjunction is true, then the first disjunction is false. The first disjunction (king or ace) is false when there is neither a king nor an ace. Either way – whichever disjunction is false – there is not an ace. The fully explicit models of the problem are accordingly:

king	¬ queen	¬ ace
¬ king	queen	¬ ace

Hence, the correct response is:

The king is more likely than the ace.

If participants took the premises to mean that one disjunction is true and the other disjunction has no definite truth value, then the other disjunction is, in effect, either true or false. The premises are then equivalent to a tautology, which supports the response that the two cards are equiprobable. A few participants make this response, but we suspect that they are guessing.

Illusory inferences come in a whole variety of forms, which we discovered by setting our computer program to search for them in the 'space' of possible inferences. Some, as we have seen, concern probabilistic reasoning. Others concern deductive reasoning. They are important because they are a robust and unexpected phenomenon predicted by the model theory.

Experiment 3

If the model theory is correct, then illusory inferences occur because reasoners fail to represent what is false, and in particular what is false within true contingencies. Experiment 3 was designed to corroborate the existence of illusory inferences about probabilities and to test this explanation.

Method

Design: The experiment was carried out at the University of Bologna, and we have translated the materials from the original Italian. Its main manipulation was to present the reasoning problems in a way that should lead individuals to consider the false contingencies. It examined three illusory problems and three matched control problems, i.e. problems that reasoners should get correct even if they fail to represent what is false. Table 7 presents these problems in abbreviated

form, together with their models, and predicted responses. In each of the six problems, the main connective was an exclusive disjunction. There were three separate groups, to which the participants were assigned at random, with equal numbers in each group. The groups differed solely in the way in which the exclusive disjunction in each problem was expressed:

True group: Only one of the two assertions is true.
False group: Only one of the two assertions is false.
True-false group: One of the two assertions is true and one of them is false.

Table 7. The three illusory problems and the three matched control problems used in Experiment 3, stated in abbreviated form, together with their predicted models and judgments. Each problem was in the form of an exclusive disjunction, i.e. one premise was true and the other premise was false. The predicted judgments for the illusory problems are wrong; those for the control problems are correct. The problems are shown with the same contents, though in fact, each problem had a different content in the experiment.

Illusory problems	Matched control problems
1. If king then ace. If not-king then ace. K A ¬K A . . . ∴ Ace more likely than king.	1'. If king then ace. If king then not-ace. K A K ¬A . . . ∴ King more likely than ace.
2. King or ace, or both. Not-king or ace, or both. K A K A ¬K A ¬K A ∴ Ace more likely than king.	2'. King and ace Not-king and ace K A ¬K A ∴ Ace more likely than king.
3. King iff not ace. King. K ¬A K ∴ King more likely than ace.	3'. Not king iff not ace. King. ¬K ¬A K ∴ King more likely than ace.

If the illusions arise because reasoners fail to represent false cases, then their performance should be better in the False group and the True-false group than in the True group. Each participant carried out all six problems presented in a different random order.

Participants: Thirty undergraduates at the University of Bologna (22 women and 8 men) volunteered to take part in the experiment, which lasted about one hour. They had not previously participated in an experiment of this sort, and they had not taken any courses in logic.

Materials: Each problem was about cards in a hand. We created three versions of each of the six problems in Table 7 by expressing them with a different lexical content based on different cards. The resulting 18 problems were then used to construct two lists, which differed only in the order of the two assertions in the individual problems. The two lists were assigned at random to the participants, ensuring that half the participants in each group had one list and half the participants had the other list. Each problem was printed on a separate card with one assertion above the other.

Procedure: The participants were tested individually. They sat at a desk in front of the experimenter, who gave them the appropriate instructions. The participants were told to imagine that they were playing cards with the experimenter, and that there were two assertions about what she had in her hand. Their task was to decide which of the two cards referred to in the assertions was more likely to be in her hand. They could decide that one card was more likely to be in the hand than the other, or that the two cards were equally likely to be in the hand. The participants were told to take as much time as they wanted to make their responses. They were not timed, because the illusory problems yield too many errors for a meaningful analysis of response times. The participants in the True group were told: Only one of the two assertions is true. The participants in the False group were told: Only one of the two assertions is false. And the participants in the True-false group were told: One of the assertions is true and one of them is false. There was one simple practice trial. If a participant failed to respond correctly, the experimenter explained the line of reasoning that lead to the correct response. The experiment proper then began. The experimenter recorded the participant's response to each problem.

Results and discussion: Table 8 presents the percentages of correct responses for the illusory inferences and their matched controls for the three groups. The illusory problems were indeed harder than the matched control inferences (25 out of the 30 participants performed worse with the illusions, and there was one tie, Sign test, $p < 0.0005$). The different ways of expressing the disjunction in the three groups also had a reliable effect, and the False-group performed better overall than the other two groups ($c2\ (2) = 18.65$, $p < 0.001$), and in particular this group performed better with the illusory problems than the other two groups ($c2\ (2) = 11.61$, $p < 0.01$). Performance in the True-false group, however, was indistinguishable from performance in the True-group.

Table 8. The percentages of correct conclusions to the three pairs of matched illusory and control problems in the three groups of Experiment 3.

	Illusory problems	Matched control problems
True group	13	58
False group	31	82
True-false group	12	62
Overall	20	67

The experiment confirmed the existence of illusory inferences – inferences that strongly suggest a conclusion that most reasoners infer, but that is totally wrong. The instruction that conveyed the exclusive disjunction using the form of words, 'Only one of the two assertions is false,' reliably improved performance, but it also improved performance on the control problems. This unexpected finding is not inconsistent with the model theory, i.e. there is no reason to suppose that such an instruction would hinder performance with control problems. The instruction, 'One of the two assertions is true and one of them is false' did not improve performance. Presumably, the participants concentrated on the 'true' part of the instruction to the detriment of the 'false' part. Even the False group, however, did not perform as well with the illusory problems as with the control problems. Hence, instructions that focus reasoners' attention on falsity are not a complete antidote to the illusions. One possibility, which we owe to Fabien Savary (personal communication), is that if reasoners are to gain a perfect understanding of the illusions, then they need to be able to do three things. First, they must understand the relevance of propositions about what is false in otherwise true states of affairs. Second, they must be able to work out what is the case when these various propositions are false. Third, they must be able to combine all of this information correctly. Our experiment merely made falsity more salient, and that in itself was not sufficient to dispel the illusions entirely.

CONCLUSIONS

The mental model theory postulates that reasoners make inferences by constructing models of the premises. Our results corroborate this theory. Experiment 1 examined deductively valid reasoning with quantified premises. It showed that reasoners find it easier to evaluate conclusions based on single models than conclusions based on multiple models. They were faster and more accurate.

The model theory extends naturally to modal reasoning. It predicts that conclusions about what is possible should be easier than conclusions about what is necessary, because possibility calls merely for a single model of the premises to support the conclusion, whereas necessity calls for

all the models of the premises to support the conclusion. This relation switches for the denial of conclusions: The denial of a possibility calls for all the models of the premises to support the conclusion, whereas the denial of necessity calls merely for a single model of the premises to support the conclusion. Experiment 2 confirmed this prediction for both speed and accuracy.

The model theory also extends to probabilistic reasoning. Our previous studies had shown that individuals reason according to the 'frequency' assumption: The more models of the premises in which an event occurs, the greater they judge the probability of the event. They also reason according to the 'equiprobability' assumption: In the absence of evidence to the contrary, they assume that models of premises represent equiprobable outcomes. The key representational assumption of the theory, however, is that individuals represent only what is true. This assumption led us – by way of a computer program implementing the theory – to a surprising prediction. There should be illusory inferences, that is, premises that strongly suggest a conclusion that is totally wrong. Experiment 3 confirmed the occurrence of illusory inferences about relative probabilities. It also provided some evidence to support the model theory's contention that the cause of the illusions is the failure to represent false contingencies. When the instructions made falsity salient by expressing an exclusive disjunction using the words:

Only one of the two assertions is false.

there was a reliable improvement in performance. The improvement appeared to occur equally with the matched control problems.

The model theory has a wider range of application than theories based on formal rules of inference (see e.g. Braine and O'Brien, 1991; Rips, 1994). As we have shown in this paper, it integrates deductive, modal, and probabilistic reasoning within a single framework: Conclusions are *necessary* if they hold in all the models of the premises; they are *probable* if they hold in most of the models of the premises; and they are *possible* if they hold in at least one model of the premises. The theory also applies to the informal arguments in scientific articles, newspaper editorials, and legal proceedings (see Shaw, 1996). Formal rule theories, however, have currently been formulated in psychology to apply only to deductive reasoning, and to limited sorts of modal inference (see Osherson, 1974-5, who specifies some formal rules for deriving modal conclusions from modal premises). What matters in resolving a controversy, however, is not the range of theories, but a crucial phenomenon. Illusory inferences are just such a phenomenon. They are robust and the model theory predicts them. But, formal rule theories neither predict nor accommodate them *post hoc*. The reason is that these theories, as currently formulated (see e.g. Braine and O'Brien, 1991; Rips, 1994), are based only on *valid* rules of inference. Hence, they cannot account for inferences in which the majority of participants draw one and the same *invalid*

conclusion. This phenomenon signals the end of the long controversy about whether human reasoning is a syntactic or semantic process.

ACKNOWLEDGEMENTS

We thank Ruth Byrne the co-author of the model theory of sentential reasoning, and Fabien Savary the co-author of the studies on illusory inferences. The research was supported in part by ARPA (CAETI) contracts N66001-94-C-6045 and N66001-95-C-8605.

REFERENCES

Bara, B., M. Bucciarelli and P. N. Johnson-Laird (1995). The development of syllogistic reasoning. *American Journal of Psychology,* **108**, 157-193.

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

Braine, M. D. S. and D. P. O'Brien (1991). A theory of If: A lexical entry, reasoning program, and pragmatic principles. *Psychological Review,* **98**, 182-203.

Chellas, B. F. (1980). *Modal logic: An introduction.* Cambridge University Press, Cambridge.

Clark, H. H. (1969). Linguistic processes in deductive reasoning. *Psychological Review,* **76**, 387-404.

Hacking, I. (1975). *The emergence of probability.* Cambridge University Press, Cambridge.

Johnson-Laird, P. N. (1994). Mental models and probabilistic thinking. *Cognition,* **50**, 189-209.

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

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

Johnson-Laird, P. N., P. Legrenzi, V. Girotto, M. S. Legrenzi and J-P. Caverni (1996). *Naive probabilistic reasoning: A mental model theory.* Unpublished paper, Department of Psychology, Princeton University.

Johnson-Laird, P. N. and F. Savary (1996). Illusory inferences about probabilities. *Acta Psychologica,* **93**, 69-90

Lewis, C. I., and C. H. Langford (1932). *Symbolic logic.* Dover, New York.

Macnamara, J. (1986). A border dispute: The place of logic in psychology. Bradford Books, MIT Press, Cambridge, MA.

Osherson, D. N. (1974-6). *Logical abilities in children,* Vols. 1-4. Lawrence Erlbaum Associates, Hillsdale, NJ.

Pollock, J. (1989). *How to build a person: A prolegomenon.* Bradford Books, MIT Press, Cambridge, MA.

Rips, L. (1994). *The psychology of proof.* MIT Press, Cambridge, MA.

Robinson, J. A. (1979). *Logic: Form and function, the mechanization of deductive reasoning.* Edinburgh University Press, Edinburgh.

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

Shaw, V. F. (1996). The cognitive processes in informal reasoning. *Thinking and Reasoning*, **2**, 51-80.

Sperber, D. and D. Wilson (1986). *Relevance: Communication and cognition.* Basil Blackwell, Oxford.

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

Wason, P. C. (1959). The processing of positive and negative information. *Quarterly Journal of Experimental Psychology*, **11**, 92-107.