COGNITIVE SCIENCE Vol 22 (1) 1998, pp. 25–51 ISSN 0364-0213

A Model Theory of Modal Reasoning

VICTORIA A. BELL AND P. N. JOHNSON-LAIRD

Princeton University

This paper presents a new theory of modal reasoning, i.e. reasoning about what may or may not be the case, and what must or must not be the case. It postulates that individuals construct models of the premises in which they make explicit only what is true. A conclusion is possible if it holds in at least one model, whereas it is necessary if it holds in all the models. The theory makes three predictions, which are corroborated experimentally. First, conclusions correspond to the true, but not the false, components of possibilities. Second, there is a key interaction: it is easier to infer that a situation is possible as opposed to impossible, whereas it is easier to infer that a situation is not necessary as opposed to necessary. Third, individuals make systematic errors of omission and of commission. We contrast the theory with theories based on formal rules.

INTRODUCTION

Consider the following two inferences. First, given the following diagnosis:

The problem is in the turbine or in the governor, or both.

It follows that:

The problem *may* be in both the turbine and the governor.

Second, given that the following premises refer to the same one-on-one game of basketball:

If Allan is in the game then Betsy is in the game.

If Carla is in the game then David is not in the game.

It follows that:

It is possible that Betsy is in the game.

What these two examples have in common is that they are both cases of *modal* reasoning, that is, reasoning about what may or may not be the case, or what must or must not be the

Direct all correspondence to: Victoria A. Bell, Department of Psychology, Princeton University, Princeton, NJ 08544 USA; E-Mail: vabell@princeton.edu; E-Mail: phil@clarity.princeton.edu

case. Many inferences in science, the law, and everyday life, concern affirmations or denials of what is possible. Indeed, the inductive generation of hypotheses and possible explanations is a sort of modal reasoning. Logicians from Aristotle onwards have studied modal logic (see e.g., Kneale & Kneale, 1962). There are many different modal logics, and logicians have developed both axiom systems for them and corresponding accounts of their semantics (see Hughes & Cresswell, 1996). But modal reasoning has been neglected by psychologists (see Evans, Newstead, & Byrne, 1993, for a magisterial review of the literature). Osherson (1976) carried out a pioneering study to test his theory based on formal rules of inference, including rules for the two main modal operators of necessity and possibility. And there have been studies of Wason's selection task for *deontic* problems—a special case of modal reasoning that concerns permission and obligation (see e.g. Cheng & Holyoak, 1985; Cosmides, 1989). Our aim in the present paper, however, is to propose a general theory of modal reasoning based on mental models, to contrast it with another nascent approach based on formal rules of inference, and to show that the evidence supports the model theory.

The *concepts* of possibility and necessity have been investigated in children's intellectual growth. As so often, the pioneers were Jean Piaget and his colleagues (see e.g., Inhelder & Piaget, 1958). But, their work has been criticized on the grounds that modal logic is a more appropriate model than Piagetian structure (see Piéraut-Le Bonniec, 1980). Piéraut-Le Bonniec herself argued that the concept of logical necessity arises from the simultaneous consideration of cases of possibility and impossibility, and she reported experimental studies that charted children's increasing grasp of these concepts. Our concern, however, is not with the way in which modal concepts develop, but with the way in which adults make modal inferences. Our plan in what follows is to describe the mental model theory of reasoning, to show how it extends naturally to deal with modal reasoning, and to establish that some new evidence corroborates the theory. Finally, we compare the model theory with potential theories based on formal rules of inference.

THE MENTAL MODEL THEORY OF MODAL REASONING

The mental model theory postulates that reasoning is a semantic process, which depends on understanding the meaning of premises (Johnson-Laird & Byrne, 1991). When individuals understand discourse, or perceive the world, or imagine a state of affairs, then according to the theory they construct mental models of the corresponding situations. In the case of verbal reasoning, they construct models from a representation of the meaning of the assertions and, where relevant, from general knowledge. This process of building discourse models is part of normal comprehension, because models at the very least are needed in order to represent the referents of discourse (Garnham, 1987). Reasoners, however, can formulate an informative conclusion from the models of the premises, and they can assess its strength from the proportion of models of the premises in which it is true (Johnson-Laird, 1994).

A mental model is, by definition, a representation that corresponds to a possibility, and that has a structure and content that captures what is common to the different ways in which the possibility could occur. A fundamental assumption of the theory is that, in order to minimize the load on working memory, people normally reason only about what is true. This principle is subtle because it applies at two levels: individuals represent only true possibilities; and they represent only the true components of these true possibilities. This point can be clarified by comparing a truth table with a set of mental models. In logic, the meaning of an exclusive disjunction, such as:

There is a circle or there is a triangle, but not both

is defined by stating how its truth value depends on the truth values of its "atomic" propositions. Thus, the exclusive disjunction is true if *there is a circle* is true (and *there is a triangle* is false), or if *there is a triangle* is true (and *there is a circle* is false). This definition can be laid out in the form of a "truth table":

There is a circle.	There is a triangle.	There is a circle or there is a triangle, but not both.
True	True	False
True	False	True
False	True	True
False	False	False

Each row in the table shows a possible combination of the truth values of the two atomic propositions, and the resulting truth value of the assertion that is an exclusive disjunction of the two. The mental models of the disjunction, in contrast, represent only the true possibilities, and within them, they represent only their true components:

o

  △

where "o" denotes a model of the circle, "△" denotes a model of the triangle, and each row denotes a model of a separate possibility. It is important not to confuse falsity with negation. Thus, consider the following exclusive disjunction, which contains a negative proposition:

There is a *not* a circle or else there is a triangle

where both propositions cannot be true. It has the mental models:

¬o

  △

where "¬" denotes negation. The first model, ¬o, does not represent explicitly that it is false that there is a triangle in this case; and the second model, △, does not represent explicitly that it is false that there is not a circle in this case. Reasoners make a "mental footnote" to keep track of the false information, but these footnotes are soon likely to be forgotten. Johnson-Laird and Byrne (1991) used square brackets as a special notation to denote these mental footnotes, but we forego this notation here. Only fully explicit models of what is possible given the exclusive disjunction represent the false components in each model:

¬o ¬Δ

o Δ

In general, a false affirmative proposition (e.g., *there is a triangle*) is represented by a true negation, and a false negative proposition (e.g., *there is not a circle*) is represented by a true affirmative.

A conditional:

If there is a circle then there is a triangle

has an explicit model of the possibility where the antecedent and consequent are true, but individuals defer a detailed representation of the possibility that the antecedent is false, i.e., where it is false that there is a circle, which they represent in a wholly implicit model denoted here by an ellipsis:

o Δ

. . .

The implicit model is a place holder: it has no explicit content. Reasoners need to make a mental footnote that possibilities where there is a circle are exhaustively represented in the explicit model, and so a circle cannot occur in the possibilities represented by the implicit model. If reasoners retain this footnote, they can transform the implicit model into a fully explicit one. Our evidence suggests, however, that reasoners soon lose access to these footnotes, especially with slightly more complex propositions.

The reader might suppose that the theory is full of murky concepts, such as "mental footnotes," square brackets that come and go, and triple dots of uncertain meaning. And it might seem to rely on hidden assumptions that can be added or dropped as needed to account for experimental results. Such appearances are misleading. The theory assumes that individuals normally reason using mental models, but in simple cases they are able to flesh out their models to make them fully explicit. Table 1 summarizes the mental models for each of the major sentential connectives, and it also shows the fully explicit models that represent the false components of true possibilities. It represents these false cases as true negations. All the specific predictions about reasoning with sentential connectives derive from this table. Reasoners make footnotes on mental models to indicate what has been exhaustively represented, and the notion is not murky, but has been implemented in several computer programs modeling both propositional and syllogistic reasoning (see Ch. 3 of Johnson-Laird & Byrne, 1991). It is helpful in some contexts to use square brackets to represent the footnotes, but there is no need for them in the present paper, and so we forego them. The triple dots represent wholly implicit models, that is, models that serve as "place holders" representing other possibilities that as yet have no explicit content. In order to model certain tasks that go beyond straightforward deduction, such as the selection task (Wason, 1966), we have made additional assumptions. We have tried to make them explicit, and indeed they have been explicit enough for certain theorists to revise (see e.g., Evans, 1993). Likewise, in the present paper, we will make some additional assumptions in

TABLE 1
The Mental Models and the Fully Explicit Models for Five Sentential Connectives

Connectives	Mental Models		Fully Explicit Models	
A and B:	A	B	A	B
A or B, not both:	A		A	¬B
		B	¬A	B
A or B, or both:	A		A	¬B
		B	¬A	B
	A	B	A	B
If A then B:	A	B	A	B
	. . .		¬A	B
			¬A	¬B
If and only if A then B:	A	B	A	B
	. . .		¬A	¬B

Note. "¬" symbolizes negation, and "..." a wholly implicit model. A footnote on the mental models for "if" indicates that A is exhaustively represented in the explicit model; and a footnote on the mental models for "if and only if" indicates that both A and B are exhaustively represented in the explicit model.

order to explain how people generate all the possibilities consistent with a set of premises. The relation between the fully explicit models and truth tables is transparent: they correspond one-to-one with the true rows in the truth tables for connectives. Each mental model, however, corresponds to the component affirmative or negative propositions in the premise that are true in the true rows.

There is a variety of evidence supporting the model theory's account of sentential deduction (see e.g., Johnson-Laird & Byrne, 1991), and Byrne and her colleagues have shown how the theory accounts for thinking about counterfactual situations (see e.g., Byrne & Tasso, 1994; Byrne, 1996). Another recent finding is the corroboration of an unexpected consequence of the theory, which we discovered by accident from its implementation in a computer program. Certain premises have initial models that support wholly erroneous conclusions, which arise from the failure to take into account false information. Experiments have confirmed the existence of these systematic fallacies (Johnson-Laird & Savary, 1996; Johnson-Laird & Goldvarg, 1997).

Simple Modal Inferences

The model theory extends naturally to account for modal reasoning. The models of a set of premises represent the possibilities given these premises. Hence, a state of affairs is possible—it *may* happen—if it occurs in at least one model of the premises. According to the theory, reasoners represent what is true in a possibility, not what is false. As an example, consider again the assertion:

The problem is in the turbine or in the governor, or both.

Reasoners should build the following set of models of this assertion (see Table 1):

```
turbine
           governor
turbine    governor
```

where "turbine" denotes a model of a problem in the turbine, and "governor" denotes a model of a problem in the governor. No knowledge prevents the formation of the third model in the set, and this model directly yields the modal conclusion:

∴ The problem *may* be in both the turbine and the governor.

This assertion is the most informative modal conclusion—given that reasoners do not represent false components of possibilities—because it concerns both of the constituent propositions in the premise. Reasoners could also infer:

∴ The problem *may* be in the turbine

or:

∴ The problem *may* be in the governor.

But, according to the model theory, they will be most unlikely to infer spontaneously:

∴ The problem *may* be in the turbine and not in the governor.

Such a conclusion contains an explicit statement (by way of a true negation) of what is false in the possibility represented by the first model of the premises (see above). The theory accordingly predicts that inferences should tend to make explicit what is true, but not what is false, within possibilities.

An experimental study corroborated the prediction (Johnson-Laird & Savary, 1995). The experiment examined single premises based on each of the following sentential connectives:

1. A and B
2. A or B, but not both
3. A or B, or both
4. A if B
5. A only if B
6. A if and only if B
7. not both A and B
8. neither A nor B

The premises concerned letters drawn on a blackboard, as in many studies of reasoning (e.g., Braine, Reiser, & Rumain, 1984), and so the participants were told that their task was to imagine one *possible* blackboard consistent with each of the premises. The premises were presented one at a time on the visual display of a computer, and the participants typed their answers beneath the premises.

The principal result was that the participants, who had no training in logic, overwhelmingly generated those possibilities predicted by the model theory, i.e., they imagined possibilities that corresponded to the mental models in Table 1, and only 5% of their responses contained explicit negations corresponding to the false cases in the fully explicit models in Table 1. In other words, their conclusions described what was true in a particular possibility, and omitted what was false in that possibility.

A KEY INTERACTION IN MODAL REASONING

You can reason from both perceptual and verbal information, and so the mental model theory postulates that reasoning can be based on perceptual models (Marr, 1982) and on discourse models (Garnham, 1987). Suppose, for example, that you have a map of a city, such as the one in Figure 1, and someone asks you:

Is it possible to go from the church to the bank via a library?

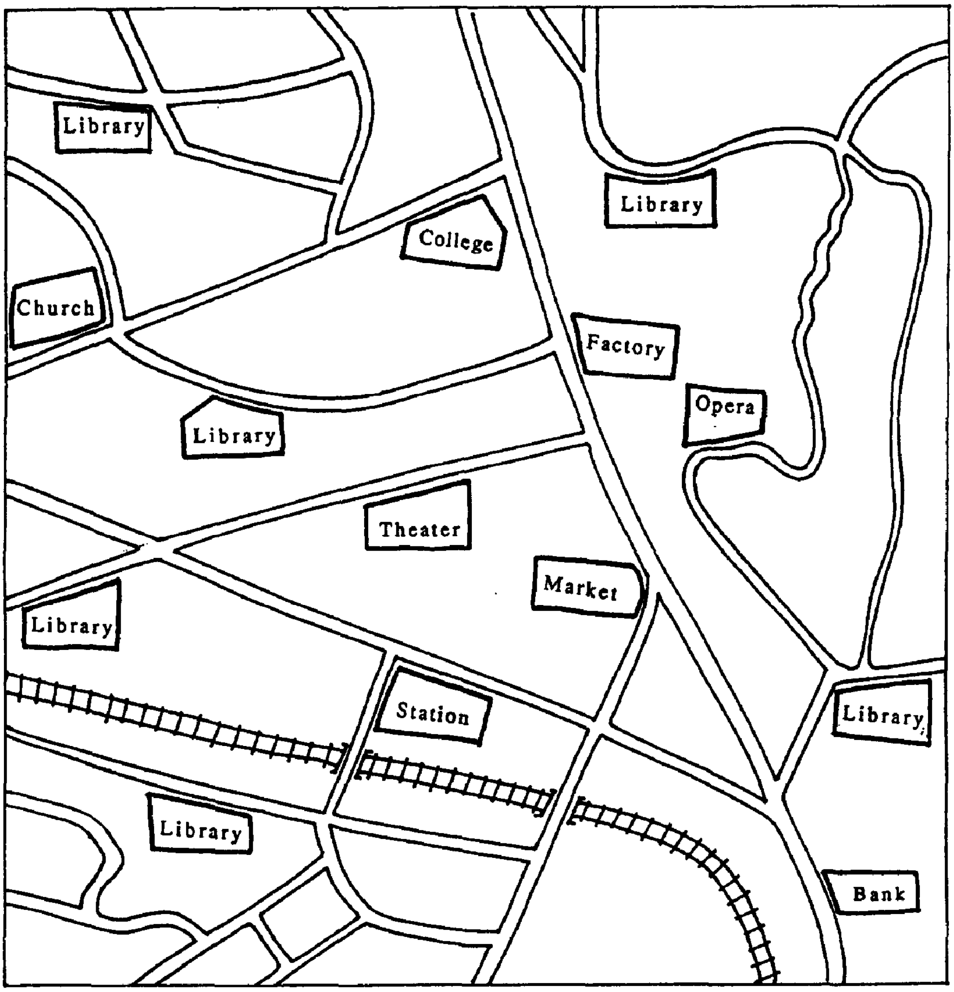

Figure 1. One of the maps used in preliminary study with the question: Is it possible to go from the church to the bank via a library? There are two direct routes from the church to the bank, and there is a library on one of them. The church and the bank were highlighted in yellow.

To answer this question truthfully, you have to examine the possible routes from the church to the bank, and, as soon as you find one that passes a library, you can respond affirmatively (the correct answer). But, you would have to examine all the routes in order to respond negatively to a question about a possibility. Suppose instead that the question is:

Is it necessary to go from the church to the bank via a library?

To answer this question truthfully, you have to examine the possible routes from the church to the bank, and, as soon as you find one that does not pass a library, you can respond negatively (the correct answer). But, you would have to examine all the routes in order to respond affirmatively to a question about a necessity. In a preliminary study, we gave 20 undergraduates eight problems of each of these four sorts. They responded "yes" to a question about a possibility (mean of 5.1 secs.) faster than they responded "no" (mean of 10.1 secs.), but they responded "no" to a question about a necessity (mean of 8.3 secs.) faster than then responded "yes" (mean of 9.3 secs.) All 20 participants followed this predicted interaction ($p = .5^{20}$, i.e., less than 1 in a million).

The interaction between modality (possibility or necessity) and polarity (affirmation or negation) is almost self-evident when individuals have to answer questions by examining maps, because we are all familiar with searching maps, diagrams, or scenes, for what is possible or necessary. Yet, the model theory predicts the same key interaction about reasoning from verbal premises. Consider the evaluation of the truth of the four sorts of assertion:

1. Affirmative possibility: a proposition, *A is possible,* is true if A is true in at least one model of the premises.
2. Negative possibility: a proposition, *A is not possible,* is true if A is false in all the models of the premises.
3. Affirmative necessity: a proposition, *A is necessary,* is true if A is true in all the models of the premises.
4. Negative necessity: a proposition, *A is not necessary,* is true if A is false in at least one model of the premises.

These evaluations bring out the intimate connection between the modal operators and the quantifiers "all" and "some," and between the model theory and the *semantics* of the modal operators in logic (see Hughes & Cresswell, 1996). They also show that a single model—an example—establishes a claim about what is possible; but all models must be counterexamples to refute such a claim. In contrast, all models must be examples to establish a claim about what is necessary; but a single model—a counterexample—suffices to refute such a claim. It follows that conclusions about what is possible should be easier to draw—faster and more accurate—than conclusions about what is *not* possible; whereas conclusions about what is *not* necessary should be easier to draw than conclusions about what is necessary. This interaction is a central prediction of the model theory, and it should apply to any sort of modal reasoning independently from the specific assumptions about the form of mental models. There were no data in the literature concerning this prediction, and so we carried out an experiment in order to test it.

Experiment 1: A Test of the Key Interaction in Verbal Reasoning

If the model theory is correct, then reasoners will construct models from their understanding of the premises, and so the key interaction should still occur. Our first attempt (in collaboration with Patrizia Tabossi) to test the prediction 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 1 to test the interaction in deontic inferences in which it was not obvious that what was possible was also necessary.

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 drawn from the premises, who play against each other. Thus, of the four players referred to in the following premises, only two can be in a game:

1. If Allan is in then Betsy is in.
   If Carla is in then David is out.
   Can Betsy be in the game?

Henceforth, we will abbreviate such problems in the following way:

1. If A is in then B is in.
   If C is in then D is out.
   Can B be in?

The first premise elicits the models (see Table 1):

A B
. . .

where "A" denotes Allan in the game, and "B" denotes Betsy in the game, and reasoners make a mental footnote that this explicit model exhausts the models in which A occurs. To answer the question, "Can B be in?", reasoners need to verify only that the explicit model above is consistent the models of the second premise. The second premise elicits the models:

C ¬D
. . .

where "¬D" denotes David out of the game, i.e., not in the game. Reasoners who go no further will judge that the model containing A and B is consistent with these models, because these two players can occur in one of the cases represented by the wholly implicit model of the second premise (denoted by the ellipsis). In fact, they will be correct, because if the second set of models is fleshed out explicitly (see Table 1), they are:

C ¬D
¬C D
¬C ¬D

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

A B ¬C ¬D

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

Must B be in?

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

B C ¬D
B ¬C D
B ¬C ¬D

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

B C ¬D
B ¬C D
A B ¬C ¬D

It follows that B 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 B is *out*. Consider the second set of models:

C ¬D
¬C D
¬C ¬D

In the first case, if B is out, then A is the only player left to be in. But, if A is in, then B should be too; and the result would be an illegal game with three players instead of two. Hence, B is in. The same argument applies *mutatis mutandis* to the second model. And, as we have seen, B and A 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 A is out then B is out.
   If C is out then D is in.

This problem has the following three fully explicit models:

¬A ¬B C D
A ¬B ¬C D
A ¬B C ¬D

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

Must B be in?

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

$\neg A \quad \neg B$

and then to establish that the second premise allows both C and D to be in. The answer to the question is accordingly, "No." In contrast, given the possible question:

Can B be in?

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 A is in then B is in

is equivalent to one of the form:

A is out and/or B is in

where "and/or" expresses an inclusive disjunction. The disjunctive equivalents of problems 1 and 2 are thus:

1'. A is out and/or B is in.
C is out and/or D is out.
2'. A is in and/or B is out.
C is in and/or D is in.

Different models are likely to be salient when the problems are expressed using disjunctions—a point to which we will return later. However, the theory still predicts the key interaction: a possibility is established by a single model and refuted only by all three models, whereas a necessity is refuted by a single model and established only by all three models.

Method. The participants served as their own controls and carried out four versions of each of eight sorts of problems—a total of 32 problems, which were presented in either one random order or its opposite. The eight sorts of problems were based on whether the premises were conditionals or disjunctions, the question was about a possibility or a necessity, and the correct answer was affirmative or negative. All the problems were consistent with three pairs of players being in the game and ruled out the other three pairs of players. The participants read the premises and then tried to answer a question, either about a possibility, e.g.:

Can Betsy be in the game?

or about a necessity:

Must Betsy be in the game?

Each problem was presented twice (with different names on the two occasions), 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

TABLE 2
The Four Problems in Experiment 1 with a Necessary Player, Together with Their Fully Explicit Models in Which the Necessary Player Is Shown in Bold

Premises	Fully Explicit Models			
If A in then B in.	A	**B**	¬C	¬D
If C in then D out.	¬A	**B**	C	¬D
	¬A	**B**	¬C	D
A out and/or B in.	¬A	**B**	C	¬D
C out and/or D out.	¬A	**B**	¬C	D
	A	**B**	¬C	¬D
If A out then B out.	**A**	¬B	C	¬D
If C in then D out.	**A**	B	¬C	¬D
	A	¬B	¬C	D
A in and/or B out	**A**	¬B	C	¬D
C out and/or D out	**A**	¬B	¬C	D
	A	B	¬C	¬D

Note. The four problems with an impossible player are the duals of these problems obtained by switching "in" for "out" and *vice versa*.

so the correct answer was negative to both questions. Table 2 presents the premises for the four problems with a necessary player and their fully explicit models. The four problems that yield an impossible player are their duals obtained by changing "in" to "out," and *vice versa*. We used "and/or" to express inclusive disjunction, because a pilot study showed that participants were confused by the tag, "or both," as in "Allan is out or Betsy is in, or both," where it was sometimes taken to mean that both players were in the game.

The eight logically distinct sorts of problems were used to construct 32 sorts of problems by assigning to them four distinct sets of 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).

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 "NO", and the H key was labeled "?" to mean "I don't know" and it 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 produce the question. They were also told to take their time to respond as "accuracy is more important than speed." The premises stayed on the screen while the question appeared underneath. After the participants made their response, the correct answer appeared below the question. 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.

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.

Results and discussion. Table 3 presents the percentages of correct responses to the four sorts of problems (affirmative possibility, negative possibility, affirmative necessity, and negative necessity), and Table 4 presents 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, $z = 1.993$, $p < 0.05$), which 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 fewer errors on negative necessities than on affirmative 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 3
The Percentages of Correct Responses to the Four Sorts of Problem in Experiment 1

	Possible Questions	Necessary Questions	Overall
"Yes" responses	91	71	81
"No" responses	65	81	73
Overall	78	76	77

Note. (n = 20).

TABLE 4
The Latencies in secs of the Correct Responses to the Four Sorts of Problem in Experiment 1

	Possible Questions		Necessary Questions		Overall	
"Yes" responses	18.0	(2.6)	25.6	(3.7)	21.8	(3.1)
"No" responses	22.3	(4.2)	22.7	(4.2)	22.5	(4.1)
Overall	20.1	(3.3)	24.1	(3.8)	22.0	(3.4)

Note. (n = 20) and in parenthese the standard errors.

The key interaction was also reliable for the response times: out of the 20 participants, 17 showed the predicted interaction in their data (Wilcoxon Test, $z = 2.912$, $p < 0.004$). The participants were faster to respond "yes" correctly to questions about possible players than to respond "no" correctly to such questions, but they were faster to respond "no" correctly to questions about necessary players than to respond "yes" correctly to such questions. Ideally, they should have been faster to respond "no" to questions about necessary players than "no" to questions about possible players. In fact, their negative responses to possible players were faster than expected, and did not differ in latency from their negative responses to necessary players. The pattern of errors, however, suggests that there may have been a trade-off between speed and accuracy for these questions.

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 not possible, but they are faster and more accurate in inferring that a player is not necessary as opposed to necessary. 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 impossible by failing to find it in any of the models of the premises. Similarly, they infer that a state of affairs is necessary by finding it in all of the models of the premises, but infer that it is not necessary by finding a counterexample to it among the models of the premises.

Errors in Modal Reasoning

The model theory makes predictions about errors in modal reasoning. In particular, the difference between mental models and fully explicit models, as shown in Table 1, predicts the sorts of errors that reasoners should make. In general, if their task is to list all the possibilities consistent with a set of premises, they should be more likely to list those that derive from the mental models of the premises than those that derive only from fully explicit models of the premises. They should be less likely to flesh out these models to make them fully explicit, because the procedure is demanding and requires them to bear in mind which propositions are exhaustively represented in the mental models. It follows that they should be susceptible to *errors of omission* of possibilities that are represented only in fully explicit models. This prediction derives solely from the standard assumptions of the model theory as reflected in Table 1.

What about the order in which people are likely to generate a list of all the possibilities consistent with the premises? They should generate the possibilities that derive from the mental models before they generate, if at all, the possibilities that derive from the fully explicit models. In order to make more refined predictions about order and about potential *errors of commission,* we implemented a computer program (in Common LISP) that was based on some additional assumptions geared to the specifics of the task. The program constructs the mental models for each premise (as shown in Table 1) in order to infer possibilities. It contains a parameter that specifies how many atomic propositions in the premises can be true, and so for one-on-one basketball games the parameter is set to two. The program builds the mental models for each premise, and then applies the same procedure to each model. It starts with whichever model has more players in the game, or, if there is no difference, with the model of the first premise. The idea here is that given, say, the following two models:

A ¬B
        C D

reasoners are more likely to generate the game, *C and D*, before they reflect on the first model. If a model has both players in the game, as in this case, it returns them as a potential response. If a model has one player in the game, as in the first of the two models above, it returns this player with each of the players in the second set of models as potential responses. If a model has neither player in the game, it returns their complement, i.e., the two players in the other set of models, as a potential response. The program checks whether a potential response violates the models of the other premise.

The prediction of errors of commission arose from two results in the literature. First, it is difficult to grasp what falsifies a conditional with a negative antecedent (Evans *et al.*, 1993, p. 50); second, it is difficult to grasp what falsifies a disjunction of one affirmative and one negative constituent (Evans *et al.*, 1993, p. 147). The program accordingly eliminates those potential responses that violate a conditional unless the conditional has a negative antecedent. It rejects those potential responses that violate a disjunction unless the disjunction is between one affirmative and one negative proposition. Finally, the program fleshes out the models explicitly and adds any additional response that emerges as a predicted error of omission. The program therefore systematically constructs the predicted conclusions and the predicted errors of commission and omission.

In order to illustrate the predictions, we will consider some typical problems. Suppose that you have to list all the possible games given the following premises:

If Abbey is in then Billy is in.
If Colin is in then Diane is in.

which, as usual, we abbreviate as:

1. If A is in then B is in.
   If C is in then D is in.

The mental models for the two conditionals are:

A B
        C D
      . . .

They yield two possible one-on-one games: *A and B*, and *C and D*. Reasoners can establish a third possibility, but only if they flesh out explicitly the mental models. They must bear in mind that A and C are exhaustively represented in the initial models, whereas B and D are not, and so they could both be in the game. The theory therefore predicts that reasoners should be more likely to draw the two responses above than this third response: *B and D*, which is a potential error of omission (see Conditional problem 1 in Table 5).

TABLE 5
The Conditional Problems and the Logically Equivalent Disjunctive Problems in Experiment 2, Together with their Initial Mental Models (on the left), and the Responses that they Yield (on the right).

	Conditional Problems			Disjunctive Problems	
1.	If A in then B in.		1'.	A out or B in.	
	If C in then D in.			C out or D in.	
	A B	A B		¬A B	B D, *B C
	C D	C D		¬C D	*A D
2.	If A out then B out.		2'.	A in or B out.	
	If C out then D out.			C in or D out.	
	¬A ¬B	C D		A ¬B	A C, *A D
	¬C ¬D	A B		C ¬D	*B C
3.	If A in then B in.		3'.	A out or B in.	
	If C out then D out.			C in or D out.	
	A B	A B		¬A B	B C, *B D
	¬C ¬D			C ¬D	*A C
4.	If A in then B in.		4'.	A out or B in.	
	If C in then D out.			C out or D out.	
	A B	A B		¬A B	B C, B D
	C ¬D	B C		¬C ¬D	A B
5.	If A out then B in.		5'.	A in or B in.	
	If C out then D out.			C in or D out.	
	¬A B	B C, *B D		A B	A B, A C
	¬C ¬D	A B		C ¬D	B C
6.	If A in then B in.		6'.	A out or B in.	
	If C out then D in.			C in or D in.	
	A B	*A B		¬A B	B C, B D
	¬C D	B D		C D	C D
7.	If A in then B out.		7'.	A out or B out.	
	If C out then D out.			C in or D out.	
	A ¬B	A C, *A D		¬A ¬B	C D
	¬C ¬D			C ¬D	A C, B C

Note. An "*" signifies a predicted error of commission.

In this domain, problems based on conditionals can be expressed in a logically equivalent form using disjunctions, but Table 1 predicts that there will be differences in performance between the two sorts of problems. To illustrate this point, consider the theory's predictions for the disjunctive version of the previous conditional problem (see Disjunctive problem 1' in Table 5):

1' A is out and/or B is in.
C is out and/or D is in.

The initial models of the first premise are:

¬A

B

¬A B

The third of these models: ¬*A and B*, is the critical one, because it concerns a pair of individuals. This analogous model for the second premise is:

¬C D

The first model suggests a pair containing B, and the second model suggests a pair containing D, so the mental models of the premises yield only one correct response: *B and D*. The responses of *B and C* and *A and D* are potential errors of commission, because they each violate a premise. When the models are fleshed out explicitly, they yield two further responses: *A and B*, and *C and D*, which are potential errors of omission.

The predictions are complicated, but they are not *ad hoc*. They follow directly from the difference between mental models and fully explicit models as shown in Table 1, from a plausible assumption about the order in which reasoners will derive conclusions from models, and from evidence in the literature about the difficulty of grasping false instances of conditionals and disjunctions. We designed Experiment 2 to test the predictions.

Experiment 2: A Test of the Predictions About Errors

The participants in the experiment had to list all the possible pairs of players for problems about one-on-one games of basketball. There are four sorts of conditionals in this domain:

If A is in then B is in.
If A is in then B is out.
If A is out then B is in.
If A is out then B is out.

Hence, there are 16 possible pairs of premises, but only twelve of these pairs yield three pairs of players who are in the game, and three pairs of players who are out of the game. The experiment therefore examined these twelve pairs. It also examined the twelve logically equivalent problems based on disjunctions. Five pairs of the twelve problems of each sort differ merely in the order of the two premises. For the purposes of prediction, the program implies that we can ignore the order of the premises, and so Table 5 presents the seven logically distinct problems, together with their mental models, the responses that they yield, and the predicted errors of commission.

Method. The participants in the experiment served as their own controls, and listed the possible pairs of players in one-on-one basketball games for 24 problems, which were presented in one random order or its reverse. Each problem concerned four different individuals. The problems were based on the twelve conditional pairs of premises that allowed three pairs of players to be in the game, and the twelve equivalent disjunctive problems (see Table 5 for the seven logically distinct problems of both sorts).

We selected 96 two syllable common names, each of five letters, and assigned unique sets of four of them at random to the 24 problems so that no name occurred more than once in a problem. The problems were presented on eight pages with three problems to a page. The participants were tested individually. They were told that the problems concerned one-on-one basketball games, and that they had to generate all the possible games for the 4 players mentioned in each pair of premises. They were allowed to take as much time as they wanted.

TABLE 6
The Percentages of Correct Responses to the Conditional Problems and the Logically Equivalent Disjunctive Problems in Experiment 2

	Conditional Problems			Disjunctive Problems	
1.	If A in then B in. If C in then D in.		1'.	A out or B in. C out or D in.	
	A B	**100**		**B D**	**100**
	C D	**100**		A B	74
	B D	82		C D	74
2.	If A out then B out. If C out then D out.		2'.	A in or B out. C in or D out.	
	C D	**100**		**A C**	**100**
	A B	**100**		A B	80
	A C	79		C D	70
3.	If A in then B in. If C out then D out.		3'.	A out or B in. C in or D out.	
	A B	**100**		**B C**	**98**
	B C	84		A B	70
	C D	75		C D	63
4.	If A in then B in. If C in then D out.		4'.	A out or B in. C out or D out.	
	A B	**100**		**B C**	**100**
	B C	**91**		**B D**	**96**
	B D	79		**A B**	**76**
5.	If A out then B in. If C out then D out.		5'.	A in or B in. C in or D out.	
	B C	**86**		**A B**	**86**
	A B	**86**		**A C**	**100**
	A C	75		**B C**	**98**
6.	If A in then B in. If C out then D in.		6'.	A out or B in. C in or D in.	
	B D	**88**		**C D**	**87**
	B C	75		**B C**	**100**
	C D	67		**B D**	**93**
7.	If A in then B out. If C out then D out.		7'.	A out or B out. C in or D out.	
	A C	**93**		**A C**	**100**
	B C	73		**B C**	**95**
	C D	82		**C D**	**71**

Note. (*n* = 22). Percentages to pairs in bold are those derived from the initial models.

Twenty four undergraduates students from Princeton University were recruited from an introductory psychology class and given class credit for their participation. None of the students had any training in logic.

Results and Discussion. Two participants in the experiment adopted bizarre strategies: one listed all 6 pairs of players as possibly in the game for most of the disjunctive problems, and the other participant generated an initial response and then almost invariably its dual. We therefore rejected their data (though it had no effect on the significance of the following statistical analyses). There was no reliable effect of the order of premises on performance, and so we collapsed the results from the twelve problems of each sort into the seven logically distinct problems. Table 6 presents the percentages of correct responses to these conditional problems and to their logically equivalent disjunctive problems. The percentages in bold in the table are those based on the initially explicit models. Overall, there were 95% of correct responses depending on the mental models, but only 76% of correct responses depending on fleshing out the models explicitly (the potential errors of omission). As the model theory predicted, this difference was reliable both for participants (15 participants followed the predicted pattern, 2 participants contravened it, and there were 5 ties, Wilcoxon's test, $z = 3.292$, $p < .001$) and for materials (all twelve of the conditional problems, and all four of the relevant disjunctive problems, followed the predicted pattern, $p = .5^{16}$, i.e., less than 1 in 60 thousand).

Overall, the percentage of predicted errors of commission was 13% in comparison to 3% of errors of commission that were not predicted. This difference was also reliable both for participants: 10 participants fit the predicted pattern, 1 participant contravened it, and 11 were ties, (Wilcoxon, $z = 2.806$, $p < .005$) and for materials: all 10 of the relevant problems—six based on conditionals, and four based on disjunctions fit the predicted pattern ($p = .5^{10}$, i.e., $p < .001$).

The computer program implementing the model theory, which we described earlier, predicts the order in which individuals should generate possibilities (see Table 5). We assessed the correlation between the order in which the participants generated their responses and this predicted order. We excluded errors of commission, and scored each protocol using Kendall's partial rank-order statistic P (Kendall & Gibbons, 1990) to allow for the fact that not every participant produced all three correct responses to a problem. Overall, 69% of the conditional trials followed the predicted trend and 71% of the disjunctive trials followed the predicted trend (21 out of the 22 participants fit the predicted trend on the majority of trials, and there was one tie, Wilcoxon test, $z = 4.018$, $p < .0002$; 8 disjunctive problems do not have potential errors of omission, but all the remaining 16 problems yielded data that fit the trend on the majority of trials, Wilcoxon, $z = 3.516$, $p < .0005$).

The fine-grained predictions of the model theory are borne out by the results. The responses that emerge at once from the mental models of the premises were drawn more often than those that call for the models to be fleshed out explicitly. This difference was highly reliable. Likewise, the theory predicts that the mental models would sometimes lead the participants into errors of commission, especially when a response was incompatible with a conditional with a negated antecedent or with a disjunction of one negative and one affirmative proposition. This prediction was also confirmed. Each conditional problem was expressed by a logically equivalent disjunction. As the theory predicted, however, the pattern of responses depended on how a problem was framed. What is salient can differ strikingly from one version to another (particularly in the case of problems 1 through 3 in Table 6). For example, with conditional problem 1 the two responses: *A and B* and *C and D*, were made by every participant, whereas with equivalent disjunctive problem 1, only the response: *B and D* was made by every participant. In general, the same information elicited very different responses depending on whether the premises were conditionals or disjunctions.

Is there some simple alternative explanation of the results? The only differences among the conditionals and the disjunctions concern who is in the game and who is out of the game. The term "out" is implicitly negative (cf. Clark, 1969), and so we examined whether the number of such assertions in the premises predicted the difficulty of the problems. There was no such relation. The percentages of absolutely correct responses were as follows in terms of the numbers of propositions about who was out:

73% for problems with no "out" propositions
69% for problems with one "out" proposition
62% for problems with two "out" propositions
63% for problems with three "out" propositions
75% for problems with four "out" propositions

There was no significant difference in performance across these conditions (Friedman test, $Xr^2 = 9.1$, $p > .05$). Although negation contributed to the predicted difficulty of certain problems, its role is highly specific: individuals have some problem in grasping the truth conditions of conditionals with negated antecedents and of disjunctions of one affirmative and one negative proposition. A more refined procedure may reveal other effects of negation, but the general pattern of results is not open to alternative interpretation in terms of negation alone.

GENERAL DISCUSSION

The model theory of modal reasoning postulates that each model of a set of premises represents a possibility. In alethic modal reasoning, it represents what is true given the content and structure of the model in all the different ways in which the possibility might be realized. In deontic modal reasoning, it represents what is permissible given the content and structure of the model. The theory makes three general predictions about inferential performance. The first prediction is that conclusions about possibilities should make explicit what is true, not what is false, given the premises. The prediction has been corroborated experimentally (Johnson-Laird & Savary, 1995).

The second prediction is of a key interaction in modal reasoning: reasoners should be faster and more accurate in establishing a possibility than in refuting it, whereas they should be faster and more accurate in refuting a necessity than in establishing it. The prediction derives directly from the use of models: a single example establishes a possibility, and only all models can refute it; whereas all models establish a necessity, and a single model refutes it. With maps and diagrams, it is hard to imagine that reasoners would use any other method than the one proposed by the model theory. A more striking result is accordingly the corroboration of the key interaction in Experiment 1. The participants were given verbal premises about one-on-one games of basketball, e.g.:

1. If Allan is in then Betsy is in.
   If Carla is in then David is out.

and its "dual":

2. If Allan is out then Betsy is out.
   If Carla is out then David is in.

In fact, no participant encountered problems about the same players, which we use here purely for illustrative purposes. After such premises, they were asked a question either about a possibility:

> Can Betsy be in the game?

or else about a necessity:

> Must Betsy be in the game?

They were faster and more accurate to answer affirmatively the question about the possibility than to answer it negatively, but faster and more accurate to answer negatively the question about the necessity than to answer it affirmatively.

Skeptics might argue that the interaction can be explained in terms of a response bias. In particular, individuals might be biased to respond "yes" to questions about possibility, but to respond "no" to questions about necessity. There are three grounds for resisting this putative explanation. First, our initial attempt to test the interaction failed (Patrizia Tabossi, personal communication), probably because the participants had to respond that something was "possible" when, in fact, it was obviously necessary. This failure showed that there is no general response bias favoring affirmative evaluations of possibilities and negative evaluations of necessities. Second, a simple response bias hardly conforms with the relatively long latencies of response in Experiment 1 (an overall mean of 22.0 secs), which comport rather better with the hypothesis that the participants were genuinely trying to work out the correct answers. Third, in a more recent unpublished study of informal reasoning, we have shown that individuals envisage possibilities for themselves rather faster than they envisage necessities. They were given a supposition, such as:

> Suppose all crimes were punishable by death

and asked to generate either a possible consequence of the supposition or else a necessary consequence. As the model theory predicts, they were reliably faster to generate possible consequences (mean 19.9 secs.) than necessary consequences (mean 22.3 secs.; Wilcoxon test, $z = 2.128$, $p < .04$). Because the content of the consequences was up to the participants to devise, this result cannot be explained in terms of a simple response bias.

The third prediction of the model theory is that reasoners should be more likely to envisage possibilities that correspond to mental models than possibilities that occur only in fully explicit models. Experiment 2 corroborated this prediction: errors of omission corresponded to fully explicit models. The experiment also showed that reasoners tend to envisage possibilities in the order predicted by the theory, i.e., they construct first those from the mental models of the premises, focusing on the first premise unless only the second premise yields a model of two players in the game, and then they construct those that depend on fleshing out the models explicitly.

The experiment examined some more fine-grained predictions that were made by a computer program based on assumptions about the specifics of the task and on results in the literature. The experiment also corroborated these predictions. Granted the difficulty of envisaging what is false, it showed that in certain cases reasoners envisage possibilities that are, in fact, impossible, i.e., they make errors of commission (see also Johnson-Laird & Barres, 1994). For example, given the premises (problem 6 in Table 5):

If Abbey is in then Billy is in.
If Colin is out then Diane is in.

then reasoners are likely to envisage that the following model is possible:

A B

because it is a true instance of the first premise. If they flesh out their models of the premises explicitly they may realize that in this case Colin does not play, and so, according to the second premise, Diane should play. But, it is relatively difficult to grasp the truth conditions of conditionals with negated antecedents—in this case the implicit negative that Colin is out, and the participants in Experiment 2 were more likely to make this sort of error of commission than other such errors.

The present theory places considerable importance on the construction of counterexamples, i.e., models in which a conclusion fails to hold. Not all proponents of mental models agree about this role for counterexamples. Polk and Newell (1995), for example, have advanced a mental model theory of syllogistic reasoning differing from the one formulated by Johnson-Laird and Byrne (1991), but giving a better account of individual differences in ability. Polk and Newell argue that reasoning depends primarily on linguistic processes, and that counterexamples seem to play little role in any sort of reasoning. We agree that verbal comprehension is crucial in reasoning, because it yields the set of mental models. But, the power of model-based reasoning derives from the fact that a model can refute a conclusion. Consider an invalid conclusion that is consistent with the premises, but that does not follow from them. A single model of the premises can refute such a conclusion. Our collaborator Monica Bucciarelli has demonstrated in an unpublished study that logically-untrained individuals are able to construct external models of premises in order to refute invalid conclusions. The results of Experiment 1 likewise corroborated the hypothesis that individuals base their negative responses to questions of the form, "Is X necessary?", on the construction of mental models of the premises in which X does not occur. We conclude that counterexamples *are* crucial, but that this role is best demonstrated by tasks that call for the refutation of conclusions.

Is there an alternative explanation for our results and, in particular, an explanation based on formal rules of inference? As Rips (1994) has shown, formal rules can be used as a general-purpose programming language. And in this sense, formal rules are almost irrefutable, because they can be used to simulate any computable theory, including the mental model theory. Indeed, the computer program implementing the model theory of modal reasoning depends on purely formal rules, because computer programs, at present, do not really understand anything. Our concern is accordingly not with formal rules in general, but with Osherson's (1976) system for modal reasoning, and with other current formal theories of reasoning, which are based on "natural deduction" rules of inference (e.g., Rips, 1994; Braine & O"Brien, 1991). The controversy between them and mental models—as all parties agree—is open to empirical resolution. We need to consider extensions to the formal rule theories, however, because none of them is powerful enough for the inferences in the present studies.

Osherson's ground-breaking theory contains formal rules of inference for the modal operators, such as:

Necessarily (A and B)
∴ Necessarily A and necessarily B

and:

Possibly A or possibly B, or both.
∴ Possibly (A or B, or both).

But, as he points out, his set of rules is incomplete because certain valid inferences cannot be proved with them. As an example, consider an inference in Johnson-Laird and Savary's (1995) study:

There is an "A" on the blackboard or there is a "B," or both.
∴ It is possible that there is both an "A" and a "B" on the blackboard.

This conclusion was the most frequent one drawn by the participants. Its formal derivation in modal logic is too complex to be very plausible psychologically, and indeed it cannot be proved in Osherson's system. The complexity arises because there cannot be a formal rule of the form:

A or B, or both.
∴ It is possible that both A and B.

Such a rule would imply that a self-contradiction was possible if A implied not B (Geoffrey Keene, personal communication). Hence, a proof calls for the further assumption that A does not necessarily imply the negation of B. Even if we extend Osherson's system so that it could prove this theorem, the system would make no use of examples or counterexamples. Hence, it is unlikely to account for the key interaction, which depends on the contrast between a single model (an example or counterexample) and sets of models as a whole.

The other current theories based on formal rules do not deal with modal reasoning (e.g., Rips, 1994; Braine & O'Brien, 1991). One way in which they could be extended to make inferences about a conclusion such as *A is possible* would be by using the rules to try to prove the negation of A. If not-A cannot be proved, then it follows that A is possible; if not-A can be proved, then it follows that A is not possible. Likewise, if A can be proved, then it follows that A is necessary; and if A cannot be proved, then it follows that A is not necessary. The drawbacks with this idea are two-fold. On the one hand, as Jonathan Evans (personal communication) has remarked, the idea is psychologically implausible—you do not normally seem to prove a possibility by failing to find a proof of its negation. On the other hand, the account predicts a different interaction to the model theory's key interaction. It should be easier to prove a conclusion than to fail to prove it, because failure calls for an exhaustive search through all possible formal derivations. It follows that a negative possibility should be easier to prove than an affirmative possibility, and that an affirmative necessity should be easier to prove than a negative necessity. Experiment 1 corroborated the model theory's interaction rather than this one.

Which of the many modal logics corresponds to everyday modal reasoning? In our view, the answer is likely to be: "none of the above." One disparity is that logic distinguishes actual situations, possibilities, necessities, and their negations, whereas ordinary language distinguishes actual situations, real possibilities, counterfactual situations, necessities, and their respective negations. Real possibilities, such as, at the time of writing:

Ross Perot wins the election in 2000

could happen—no matter how remote their likelihood—given the actual world. Counterfactual situations were once real possibilities, but are so no longer because they did not occur:

Bob Dole won the election in 1996.

There are several other divergences between logic and life. In logic, the necessity of *p* implies *p*. In life, claims about necessity normally reflect a process of inference and therefore are usually weaker than factual claims. For example, the assertion:

He must have won the election

is weaker in force than:

He has won the election.

One feature that distinguishes different modal logics is the consequence of iterating modal operators. Some logics distinguish between different iterations, such as *possibly p* and *possibly (possibly p)*, whereas other logics introduce axioms that allow iterations to be simplified so that the only modalities that are logically distinct are necessity, possibility, and factuality, together with their negations. In everyday life, however, the situation appears to incommensurable with either approach. An assertion, such as:

It is possible that he may have won the election

which contains two expressions of possibility, is not *logically* distinct from:

He may have won the election.

Yet, the first assertion seems to imply that the possibility is more remote than the one expressed by the second assertion. The second assertion calls for the construction of a set of models containing the event in question:

w

...

where "w" denotes a model of his having won the election. The first expression, however, seems to suggest that there is a set of alternative sets of possibilities within which the event occurs:

... ... w

...

Perhaps the decisive evidence against extensions of formal rule theories in psychology comes from a study of fallacies in modal reasoning (Johnson-Laird & Goldvarg, 1997). The model theory postulates that individuals normally reason about the truth. This assumption has a surprising consequence. As a computer program implementing the theory revealed, it implies that individuals are programmed to reason in a systematically fallacious way. Consider, for instance, the following modal problem:

Only one of the following premises is true about a particular hand of cards:

There is a king in the hand or there is an ace, or both.

There is a queen in the hand or there is an ace, or both.

There is a jack in the hand or there is a 10, or both.

Is it possible that there is an ace in the hand?

According to the model theory, reasoners will consider the possibilities given the truth of each of the three premises. For the first premise, they consider three models:

king

ace

king ace

These models suggest that an ace is possible. The second premise also suggests that an ace is possible. Hence, individuals should respond, "yes." Nearly all logically-naive individuals drew this conclusion, i.e., 99% of responses in two separate experiments (Johnson-Laird & Goldvarg, 1997). Yet, it is a fallacy that an ace is possible, because if there were an ace in the hand, then two of the premises would be true, contrary to the rubric that only one of them is true. The same strategy, however, yielded a correct response to a control problem in which only one premise refers to an ace. The participants also succumbed to fallacies of impossibility that elicited a predicted "no" response. Their confidence in their conclusions did not differ between the fallacies and the control problems. They were highly confident in both, and indeed the fallacies are so compelling that they have the flavor of cognitive illusions.

If the fallacies had not occurred, then the model theory would have been disconfirmed. In contrast, they are counterexamples to current formal rule theories. Such theories contain only logically impeccable rules, and so the only mistakes they allow are mistakes in applying a rule. Such mistakes, as Rips (1994) points out, should occur arbitrarily and have diverse results. It follows that these theories cannot begin to explain either the fallacies or the systematic errors of commission and omission that we observed in Experiment 2.

The model theory offers a unified account of reasoning about what is necessary, probable, and possible. A conclusion is necessary if it holds in all the models of the premises; and it is probable if it holds in most models of the premises. And, as our present results have shown, a conclusion is possible if it holds in at least one model of the premises, and not possible if it fails to hold in any of the models of the premises.

Acknowledgments: We thank Ruth Byrne the co-author of the model theory of sentential reasoning, and Fabien Savary and Yevgeniya Goldvarg the co-authors of the studies of systematic fallacies. We thank other colleagues for helpful criticisms and suggestions: Patricia Barres, Kyung-Soo Do, Zachary Estes, Bonnie Meyer, Mary Newsome, Lisa Torreano, Isabelle Vadeboncoeur, and Yingrui Yang. We are also grateful to Jonathan Evans, Keith Holyoak, Daniel L. Schwartz, an anonymous reviewer, and Jim Greeno for their helpful comments on an earlier version of the paper. The research was supported in part by ARPA (CAETI) contracts N66001-94-C-6045 and N66001-95-C-8605.

REFERENCES

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

Braine, M. D. S., Reiser, B. J., & Rumain, B. (1984). *Some empirical justification for a theory of natural propositional logic. The psychology of learning and motivation,* Vol. 18. New York: Academic Press.

Byrne, R. M. J. (1996). A model theory of imaginary thinking. In Oakhill, J., & Garnham, A. (Eds.), *Mental models in cognitive science.* Hove: Erlbaum (UK). Taylor & Francis. pp. 155–174.

Byrne, R. M. J., & Tasso, A. (1994). Counterfactual reasoning: Inferences from hypothetical conditionals. In Ram, A., & Eiselt, K. (Eds.), *Proceedings of the sixteenth annual conference of the cognitive science society.* Hillsdale, NJ: Lawrence Erlbaum Associates. pp. 124–129.

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

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

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

Evans, J. St. B. T. (1993). The mental model theory of conditional reasoning: Critical appraisal and revision. *Cognition, 48,* 1–20.

Evans, J. St. B. T., Newstead, S. E., & Byrne, R. M. J. (1993). *Human reasoning: The psychology of deduction.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Garnham, A. (1987). *Mental models as representations of discourse and text.* Chichester: Ellis Horwood.

Hughes, G. E., & Cresswell, M. J. (1996). *A new introduction to modal logic.* London and New York: Routledge.

Inhelder, B., & Piaget, J. (1958). *The growth of logical thinking from childhood to adolescence.* London: Routledge & Kegan Paul.

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

Johnson-Laird, P. N. and Barres, P. E. (1994). When "'or' means "'and': A study in mental models. *Proceedings of the sixteenth annual conference of the cognitive science society.* Hillsdale, NJ: Lawrence Erlbaum Associates, pp. 475–478.

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

Johnson-Laird, P. N., & Goldvarg, Y. (1997). How to make the possible seem possible. In Shafto, M. G., & Langley, P. (Eds.), *Proceedings of the nineteenth annual conference of the cognitive science society.* Mahwah, NJ: Lawrence Erlbaum Associates. pp. 354–357.

Johnson-Laird, P. N., & Savary, F. (1995). How to make the impossible seem probable. In Moore, J. D., & Lehman, J. F. (Eds.), *Proceedings of the seventeenth annual conference of the cognitive science society*. Mahwah, NJ: Lawrence Erlbaum Associates, pp. 381–384.

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

Kendall, M. G., & Gibbons, J. D. (1990). *Rank correlation methods*, 5th Edition, London: Edward Arnold.

Kneale, W., & Kneale, M. (1962). *The development of logic*. Oxford: Oxford University Press.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. San Francisco: W. H. Freeman.

Osherson, D. N. (1976). *Logical abilities in children, Vol. 4: Reasoning and concepts*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Piéraut-Le Bonniec, G. (1980). *The development of modal reasoning: Genesis of necessity and possibility notions*. New York: Academic Press.

Polk, T. A., & Newell, A. (1995). Deduction as verbal reasoning. *Psychological Review, 102*, 533–566.

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

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

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