A Model Theory of Modal Reasoning

Philip N. Johnson-Laird
Department of Psychology
Princeton University
Green Hall
Princeton, NJ 08544
phil@clarity.princeton.edu

Victoria Bell
Department of Psychology
Princeton University
Green Hall
Princeton, NJ 08544
vabell@princeton.edu

Abstract

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. A conclusion is possible if it holds in at least one mental model, whereas it is necessary if it holds in all the models. The theory makes a crucial prediction, which we corroborated experimentally. 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.

Introduction

Consider the following inference: Given the diagnosis:

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

it follows that:

The flaw may be in both the turbine and the governor.

This is an example of modal reasoning, that is, reasoning about what may or may not be the case, or what must or must not be the case. Logicians from Aristotle onwards have studied modal logic (Hughes & Cresswell, 1996); but it has been neglected by psychologists (though cf. Osherson, 1976). Our aim in the present paper is to make good the omission. We will propose a theory of modal reasoning, contrast it with another nascent approach to the topic, and show that the evidence supports our theory.

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. In the case of verbal reasoning, models are constructed from a representation of the linguistic meaning of the assertions and, where relevant, from general knowledge. Reasoners formulate an informative conclusion from the models of the premises, and they 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 set of situations, and that has a structure and content that captures what is common to these situations. A fundamental assumption of the theory is that, in order to minimize the load on working memory, people normally take into account only 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.

Consider the following exclusive disjunction, for example:

There is a not a circle or else there is a triangle, but both propositions cannot be true.

The mental models of the disjunction represent only the true possibilities, and within them, they represent only their true components:

¬o
Δ

where '¬' denotes negation, 'o' denotes a model of the circle, 'Δ' denotes a model of the triangle, and each row denotes a model of a separate possibility. Hence, the first model 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 this false information, but these footnotes are soon likely to be forgotten. Indeed, the failure to cope with falsity gives rise to illusory inferences about modal conclusions, i.e. inferences that nearly everyone makes, but that are wrong (Johnson-Laird & Goldvarg, 1997). Only fully explicit models of what is possible given the exclusive disjunction represent the false components in each model:

¬o ¬Δ
o Δ

where 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.

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 propositions. It represents these false cases as true negations. 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. The mental models are simpler: they correspond to the component affirmative or negative propositions in the premise that are true in the true rows.

The model theory accounts for modal reasoning. Each model represents a set of possibilities that have in common the structure and content of the model. 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:

Table 1: The mental models and the fully explicit models for five sentential connectives: '¬' denotes negation and '. . .' denotes a wholly implicit model.

Connective	Mental models		Fully Explicit models	
A and B:	A	B	A	B
A or else B:	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

The flaw is in the turbine or in the governor, or both. Reasoners should build the following set of models of this assertion:

turbine
governor
turbine governor

where 'turbine' denotes a model of a flaw in the turbine, and 'governor' denotes a model of a flaw in the governor. The third model yields the modal conclusion:

The flaw 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.

Suppose that there is a flaw in only two of the following components of a complex system:

If there's a flaw in the governor then there's a flaw in the turbine.

If there's a flaw in the dynamo then there is not a flaw in the transformer.

The first assertion has the following mental models (see Table 1):

governor turbine
. . .

where the ellipsis denotes a wholly implicit model, and reasoners should make a mental footnote that the explicit model exhausts the models in which the flaw is in the governor. Likewise, the second assertion has the models:

dynamo ¬ transformer
. . .

Since the flaw can be in only two of the four components, the premises yield the following three possibilities:

governor turbine
turbine dynamo
turbine transformer

And these models support the modal conclusion:

∴ The flaw must be in the turbine.

because it holds in all the models of the premises. They also support the modal conclusion:

∴ The flaw may be in the transformer.

because it holds in at least one model of the premises.

The previous example illustrates an important principle. The model theory predicts a key interaction between the modality of a conclusion (whether it concerns a possibility or a necessity) and the polarity of the conclusion (whether it is affirmative or negative). A single model an example - establishes a claim about what is possible, e.g. the flaw may be in the transformer; but all models must be counterexamples to refute such a claim, e.g. the flaw cannot be in both the governor and the dynamo. In contrast, all models must be examples to establish a claim about what is necessary, e.g. the flaw must be in the turbine; but a single model a counterexample suffices to refute such a claim, e.g. the flaw need not be in the governor. These principles are, of course, closely related to the semantic interpretations invoked by logicians for the four sorts of modal propositions, which relate the meaning of necessity to the universal quantifier and the meaning of possibility to the existential quantifier (Hughes & Cresswell, 1996). 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 the central prediction of the model theory of modal reasoning. There were no data in the literature concerning this prediction, and so we carried out several experiments in order to test it.

Experiment 1: The Interaction in Visual Reasoning

The first experiment examined the key interaction in the domain of visual reasoning. The participants in the experiment were asked questions about buildings along various routes in maps. The questions concerned either possibilities, e.g.: Is it possible to go from the church to the bank via a hotel? or else necessities, e.g.: Is it necessary to go from the church to the bank via a hotel? The correct answers to the questions, depending on the map, were either 'yes' or else 'no'. The predicted interaction is that the participants' affirmative answers should be faster than their negative answers for what is possible, whereas their negative answers should be faster than their affirmative answer for what is necessary. Only one route is needed to affirm a possibility (or to deny a necessity), whereas all routes are needed to deny a possibility (or to affirm a necessity).

Twenty Princeton and Rutgers students, with no training in logic, served as their own controls. They carried out four problems in each of eight conditions based on the modality of the question ('possible' or 'necessary'), the polarity of its correct answer ('yes' or 'no'), and the number of possible routes from the two landmarks referred to in the question (two or four). Each participant carried out the 32 trials in one of two different random orders of 32 maps.

Table 2 presents the latencies of the correct responses.

Table 2: The latencies (in seconds) of the correct responses to the four sorts of problems in Experiment 1 (n = 20).

	Possible questions	Necessary questions	Overall
Yes' responses	5.1	9.3	7.2
'No' responses	10.1	8.3	9.2
Overall	7.6	8.8	8.2

Overall, the participants were faster to respond correctly to questions about possibility (mean 7.6 seconds) than to questions about necessity (mean 8.8 seconds; Wilcoxon test, $z = 2.65$, $p < .005$). Likewise, they were faster to respond affirmatively (mean 7.2 seconds) than to respond negatively (mean 9.2 seconds; Wilcoxon test, $z = 3.66$, $p < .0005$). But, there was no significant difference in response times to the two-route and four-route problems. More important than these effects, however, is the key interaction between modality and polarity, which is evident in the table. All 20 participants followed the predicted interaction ($p = .5^{20}$, i.e. less than 1 in a million), that is, they were faster to respond affirmatively than negatively to questions about possibilities, whereas they were faster to respond negatively than affirmatively to questions about necessities.

The interaction between modality and polarity is almost self-evident when individuals answer questions from maps. Given a question about a possibility, they can answer affirmatively as soon as they have found a route passing a target, whereas they can answer negatively only after they have checked all routes. Conversely, given a question about a necessity, they can answer negatively as soon as they have found a route that does not pass a target, whereas they can answer affirmatively only after they have checked all the routes.

The corroboration of the predicted interaction shows that at least in one domain visual reasoning - an approach to modal reasoning based on mental models is highly plausible. A more stringent test of the model theory's key interaction is provided by reasoning from verbal premises. If the theory is correct, then reasoners will construct models from their understanding of the premises, and so the interaction should still occur. We tested this prediction in Experiment 2.

Experiment 2: The Interaction in Verbal Reasoning

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, only two can be in the game:

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

The first premise elicits the models:

A B

. .

where 'A' denotes a model of Allan in the game, and 'B' denotes a model of Betsy in the game, and reasoners make a mental footnote that the explicit model exhausts the models in which A occurs. To answer the question, 'Can Betsy be in?', reasoners need to verify only that the explicit model above is consistent with the second premise, i.e. it is a member of the set of models of both premises. The second premise elicits the models:

C ¬D

. . .

where '¬D' denotes a model of 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, they are:

C ¬D
¬C D
¬C ¬D

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

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

Must Betsy be in the game?

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 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 Allan is out then Betsy is out.
   If Carla is out then David 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:

¬A ¬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 Allan is in then Betsy is in.

is equivalent to one of the form:

Allan is out and/or Betsy 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. 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.

Twenty Princeton undergraduates untrained in logic served as their own controls for a total of 32 problems, which were presented in either one random order or its opposite. They carried out four versions of each of eight sorts of problems which 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. 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 equivalent player.

Table 3 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 seconds). 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).

Table 3: The percentages of correct responses to the four sorts of problems in Experiment 2 (n = 20) and in parentheses the latencies of the correct responses in seconds.

	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 corroborated by the pattern of correct responses: the participants were correct more often on affirmative possibilities than on negative possibilities, but they were correct more often 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, $z = 1.826$, $p < 0.04$). 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 speed accuracy trade-off for these questions.

In general, the results bear out the model theory's prediction of a key interaction. This robust pattern is only to be expected if reasoners make inferences from verbal premises in the same way that they make inferences from maps - in both cases, they work from models of situations, either of routes (in the previous experiment) or of games (in the present experiment). They 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.

Conclusions

The model theory of modal reasoning postulates that each model of a set of premises represents a possibility, i.e. 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. The theory predicts 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. Experiment 1 corroborated the key interaction in a domain in which participants used maps to answer questions about routes, such as: 'Is it possible to go from the church to the bank via a hotel?'. 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 2 in which the participants were given verbal premises about one-on-one games of basketball.

Is there an alternative explanation for our results and, in particular, an explanation based on formal rules of inference? We cannot answer this question decisively because no existing psychological theory based on formal rules is powerful enough for the inferences in the present studies (e.g. Braine and O'Brien, 1991; Rips, 1994). Nevertheless, we are skeptical that such a theory could be developed that would provide an alternative account of the phenomena. The difficulties are twofold. First, formal theories do not distinguish between truth and falsity these concepts are semantic, not syntactic and so it is hard to see how they could accommodate the principle that individuals take into account what is true, not what is false. Second, although Osherson's (1976) pioneering theory provides guide-lines for the development of a complete formal theory, it makes no use of examples or counterexamples. Hence, it can give no account of the key interaction, which depends on the contrast between a single model (an example or counterexample) and sets of models as a whole.

The model theory provides a 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; 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, not possible if it fails to hold in any of the models of the premises, necessary if it holds in all the models of the premises, and not necessary if it fails to hold in any models of the premises.

Acknowledgments

We thank Ruth Byrne the co-author of the model theory of sentential reasoning. We are also grateful to other colleagues for helpful criticisms and suggestions: Patricia Barres, Kyung-Soo Do, Zachary Estes, Yevgeniya Goldvarg, Bonnie Meyer, Mary Newsome, Fabien Savary, Lisa Torreano, and Isabelle Vadeboncoeur.

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.

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

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

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

Johnson-Laird, P.N., & Goldvarg, Y. (1997). How to make the impossible seem possible. See this volume.

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

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.