How to Make the Impossible Seem Probable

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

Fabien Savary
Department of Psychology
Princeton University
Green Hall
Princeton, NJ 08544
fabien@phoenix.princeton.edu

Abstract

The mental model theory postulates that reasoners build models of the situations described in premises. A conclusion is possible if it occurs in at least one model; it is probable if occurs in most models; and it is necessary if it occurs in all models. The theory also postulates that reasoners represent as much information as possible in implicit models. Experiment 1 showed that, as predicted, conclusions about possible situations tend to correspond to explicit models rather than to implicit models. Experiment 2 yielded a discovery: there are illusory inferences with conclusions that seem plausible but that are in reality gross errors. In such cases, as the model theory predicts, subjects judge as the more probable of two events one that is impossible. For example, given that only one of the following two assertions is true:

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

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

subjects judge that the ace is more likely to be in the hand than the king. In fact, it is impossible for an ace to be in the hand.

Introduction

Consider the following problem:

If there is a king or a queen in the hand then there is an ace in it.

Which is more likely to be in the hand: a king or an ace?

Most people respond correctly that the ace is more likely to be in the hand than the king. Such judgments have not hitherto been studied in the psychological laboratory, and there is only one theory that purports to explain them -- the theory of mental models. This theory also predicts that certain premises should give rise to what we call illusory inferences. These yield conclusions that nearly everyone draws, that seem plausible, and yet that are egregious errors. Our plan in what follows is, first, to outline the theory of mental models; second, to describe a study that corroborates its account of inferences about what is possible; and, third, to describe the study that confirmed the existence of illusory inferences.

The Mental Model Theory of Reasoning

The theory of mental models (see e.g. Johnson-Laird and Byrne, 1991) postulates that reasoning -- deductive or inductive -- is a process in which reasoners first represent the truth conditions of premises, and then use this representation together with their semantic and general knowledge to construct mental models of the relevant situations. These models may take the form of visual images, but their critical feature is their structure. Thus, a simple conjunction:

There is a king in the hand and there is a queen in it too

calls for a single model, which we represent in the following diagram where 'K' denotes a king and 'A' denotes an ace:

K A

Likewise, the exclusive disjunction:

There is a king or there is an ace, but not both

calls for two alternative models (one for each possibility), which we represent in the following diagram:

K

A

where each line represents a separate model. The representation of explicit information is kept to a minimum so as not to overload working memory. The models of the disjunction are thus partially implicit because they do not make explicit that an ace does not occur in the first model and that a king does not occur in the second model. Reasoners thus need to make a mental 'footnote' that the first model exhausts the hands in which a king occurs and the second model exhausts the hands in which an ace occurs. (Johnson-Laird and Byrne, 1991, used square brackets to represent such a footnote, but we will forego that notation here.) The footnote, provided it is remembered, can be used to make the models wholly explicit if necessary:

K ¬A

¬K A

where '¬' denotes negation.

The same general principles underlie the initial representation of a conditional:

If there is a king then there is an ace.

Individuals grasp that the conditional means that both cards may be in the hand, which they represent in an explicit model, but they defer a detailed representation of the case where there is not a king in the hand, which they represent in a wholly implicit model denoted here by an ellipsis:

K A

. . .

Reasoners need to make a mental footnote that hands in which a king occurs are exhaustively represented in the explicit model, and so a king cannot occur in the hands represented by the implicit model. But, since hands containing an ace are not exhausted in the explicit model, they may, or may not, occur in the hands represented by the implicit model.

Because assertions may include several connectives, e.g. "A or B, and C or D", the program simulating the mental model theory is recursive, and depends ultimately on the principles for conjunction and negation. The conjunction of two sets of models (corresponding to assertions conjoined by 'and') calls for the pairwise combination of each member of one set with each member of the other set. The principles of conjunction are summarized in Table 1. Negation, in principle, calls for the construction of the complement of a set of models, but in practice is more complicated. In general, if a conclusion holds in all the models of the premises, it is necessary; if it holds in most of the models, it is probable; and if it holds in at least one model, it is possible.

Table 1: To form a conjunction of two sets of models: use the appropriate principle for each pairwise combination.

1. For two explicit models, conjoin their elements, dropping any duplicates, e.g.:

A B and B C => A B C

If one model (bearing in mind any footnote) contains an element, A, that contradicts an element, ¬A, in the other, the result is the null model (akin to the empty set). When reasoners conjoin two separate premises, they tend to drop propositions that they know categorically, e.g.:

A and A B => B, if A is categorical.

2. The conjunction of an implicit model with an explicit model is, in principle, constrained by any footnotes on the sets of models (and follows principle 1). If the footnotes have been forgotten, then the result is the explicit model, e.g.:

. . . and B C => B C

unless this model is already a member of the set of models containing the implicit model, in which case the result is the null model.

3. The conjunction of two implicit models should also be constrained by footnotes (and principle 1). If the footnotes have been forgotten, then the result is an implicit model:

. . . and . . . => . . .

unless one has already been formed from the conjunction.

Previous studies have corroborated the predictions of the model theory about valid deductions, and, in particular, they have shown that the greater the number of models that have to be constructed to draw a necessary conclusion, the harder the task is -- it takes longer and is more prone to error (see e.g. Johnson-Laird and Byrne, 1991).

Mental Models and Possibility

A conclusion describes a possible state of affairs if at least one model of the premises supports it. The theory predicts that such conclusions should correspond to explicit models of the premises rather than to implicit models. For example, the premise:

There is an 'A' on the blackboard or there is a '2', or both

has the explicit models:

A	
	2
A	2

Hence, if subjects have to describe the contents of a possible blackboard, their response should match one of these three models rather than be, say, 'A not-2', which is logically correct but which contains an item that is not in an explicit model. In the first stage of Experiment 1, we tested 26 Princeton students with such premises. We rejected three subjects because they failed to follow the instructions: they tended to describe several alternatives rather than just one. Table 2 summarizes the results. It treats the premises as though they had the same content; in fact, the subjects saw each letter-number combination only once. Most of the "other" responses merely included letters other than those mentioned in the premise. The data bore out the prediction: 17 out of 23 subjects made a majority of responses corresponding to explicit models (Sign test, with one tie, $p < .01$).

Table 2: The percentages of responses to the eight premises in the first stage of Experiment 1. The subjects (n = 23) described one possibility given the truth of the premise. Percentages in bold are for the responses predicted to be the most frequent by the model theory.

Type of premise	Responses and their percentages			
	A	B	AB	other
A if B	15	7	**56**	22
A only if B	13	11	**54**	22
A if and only if B	0	18	**54**	28
A and B	0	6	**57**	37
A or B, or both	**39**	**9**	**41**	11
A or B, not both	**28**	**44**	0	28
not both A and B	**46**	**24**	0	**30**
neither A nor B	0	2	0	**98**

In stage 2 of Experiment 1, the subjects again responded with a possible state of affairs, but each trial was based on two premises, which both had to be taken into account in the response. Table 3 summarizes the results. "Other" responses included again unmentioned letters, the letter "C" alone, negated literals, and responses naming specific numbers of letters. Once again, however, 17 out of the 23 subjects draw a majority of conclusions corresponding to explicit models (Sign test, $p < .02$).

We draw two morals from the experiment. First, logically-untrained individuals can draw correct conclusions about what is possible both from single premises and from pairs of premises. Second, as the model theory predicts, their conclusions tend to correspond to explicit models.

Illusory Inferences about Probabilities

The computer program implementing the model theory predicted the existence of a novel category of inferences that have a striking property: their initial models yield a

Table 3: The percentages of responses to the eight problems in stage 2 of Experiment 1. The subjects stated what was possible given the truth of the premises. Bold responses are those predicted by the model theory.

Type of problem	Responses and their percentages						
	A	B	AB	AC	BC	ABC	Other
1. A if and only if B C if and only if A	0	9	2	0	0	**70**	19
2. A if B C if A	0	7	0	0	0	**76**	17
3. A or B, not both C if not-A	**18**	**2**	4	4	**52**	0	20
4. A or B, or both C if not A	**2**	**5**	**28**	0	**41**	0	24
5. not both A and B C if and only if A	7	**17**	2	**54**	0	0	20
6. not both A and B C if not A	**4**	9	2	2	**57**	0	26
7. A or B, not both B or C, not both	4	**33**	0	**46**	0	0	17
8. A or B, or both B or C, or both	4	**11**	**9**	**2**	**4**	**57**	13

conclusion that is opposite to the one supported by the fully explicit models of the premises. Hence, if the theory is correct, these inferences should give rise to an illusion: nearly everyone should draw the same conclusion, it should seem obvious, and yet it will be totally wrong. In this section, we will outline the model theory's predictions and describe some corroboratory results.

Here is an example of an illusory inference:-

1. Suppose that only one of the following assertions is true about a hand of cards:

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

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

Which is more likely to be in the hand: a king or an ace?

The models of the first premise are:

K

A

K A

and the models of the second premise are:

Q

A

Q A

The assertion that only one of the two premises is true calls for an exclusive disjunction of them, and an exclusive disjunction, X or else Y, has the following initial models:

X

Y

Hence, the initial models of the premises merely include all the models above. The probability of an event is estimated on the basis of the proportion of models in which it holds (Johnson-Laird, 1994), i.e. reasoners will tend to assume that models are equiprobable. Hence, they will respond that the ace is more probable than the king. If they made no such assumption, they would conclude that the problem is indeterminate, e.g. the probability of the king alone could be greater the probabilities of all the other models summed together. Both of these responses are wrong, however.

What has gone wrong? The two disjunctions are in an exclusive disjunction, and so when one is true, the other is false. When the first disjunction is false there is neither a king nor an ace, and when the second disjunction is false there is neither a queen nor an ace. In fully explicit models, the first disjunction is combined with the negation of the second disjunction, and the second disjunction is combined with the negation of the first disjunction. Hence, the fully explicit models of the premises are:

K ¬Q ¬A

¬K Q ¬A

The king is possible, but the ace is impossible, and so the correct response is that the king is more probable.

We also used an illusory inference based on conditionals:

2. Suppose that only one of the following assertions is true about a hand of cards:

If there is a king in the hand then there is an ace.

If there is a queen in the hand then there is an ace.

Which is more likely to be in the hand: a king or an ace?

The initial models suggest that the ace is more probable than the king, but the fully explicit models show that an ace cannot occur in the hand, and so the king is actually more probable than the ace.

We gave 24 Princeton students the two illusory inferences together with two inferences that the model theory predicts will elicit the correct responses (because the initial models support the same conclusion as the fully explicit models). Each subject received the four problems in a different order, and each problem was based on a different set of cards.

The results are summarized in Table 4. The subjects were correct on 71% of the control problems, but only on 17% of the illusory inferences. 20 out of the 24 subjects were more accurate with the control inferences than with the illusory ones, and there were two ties (Sign test, $p < .001$). Overall, 21 out of the 24 subjects chose as more probable for one or both of the illusory problems a card that could not occur in the hand.

Table 4: The percentages of responses to illusory and control inferences. Correct responses are in bold.

Type of problem	Percentages of responses		
	ace	king	equi-probable
Illusory inferences:			
1. Only one assertion is true: king or ace, or both. queen or ace, or both. Which is more likely: king or ace?	75	**21**	4
2. Only one assertion is true: If king then ace. If queen then ace. Which is more likely: ace or king?	79	**13**	8
Control inferences:			
3. If king then ace. Which is more likely: king or ace?	**62**	17	21
4. If king or queen then ace. Which is more likely: ace or king?	**79**	17	4

Conclusions

The model theory was originally developed as an account of how people draw logically necessary conclusions. But, the theory also provides an obvious mechanism for reaching conclusions about what is possible or what is probable: a situation is possible if it holds in at least one model of the premises, and it is probable if it holds in most models of the premises (Johnson-Laird, 1994). The twist in these predictions is that reasoners are likely to use implicit models and to forget the mental footnotes that constrain their contents. It follows that a conclusion about what is possible should tend to correspond to an explicit model because implicit models have, by definition, no immediately available content. Our first experiment confirmed this prediction. It is worth emphasizing that the prediction is based on the models needed for deduction, that is, we did not develop a new theory to account for reasoning about possibilities. The same prediction might be derived from a theory based on formal rules of inference (e.g. Rips, 1994), but such a theory will have to introduce special 'modal' rules for dealing with possibilities.

The model theory predicted that there should be illusory inferences, and our results show that they exist. Colleagues who have succumbed to an illusion have suggested that they perhaps misinterpreted the assertion: "Only one of the following assertions is true", and that they took it to mean (1) that one assertion was true and the other was of unknown truth value, or (2) that the two assertions were in an inclusive disjunction, or (3) that the two assertions were in a conjunction. A recent unpublished study shows that none of these hypotheses is correct. The first two hypotheses are equivalent: an assertion of an unknown truth value is either true or false, and models of a disjunction, X or Y, that take the form:

X and (Y or ¬ Y)
(X or ¬X) and Y

are equivalent to those of an inclusive disjunction:

X Y
X ¬ Y
¬ X Y

In our recent study, we examined the following illusion:

Only one of the following assertions is true:
If there is a king in the hand then there is an ace.
If there isn't a king in the hand then there is an ace.

Nearly everyone judged that the hand is more likely to contain an ace than a king; in a separate condition of the experiment, they also deduced that the ace was in the hand. In fact, it is impossible for there to be ace in the hand. But, an inclusive disjunction of the two conditionals yields a tautology (there is or isn't a king in the hand, and there is or isn't an ace in the hand), and this interpretation cannot predict the illusion.

The third hypothesis -- that subjects treat the connective as a conjunction -- is also refuted by the same study. It included a control problem of the form:

Only one of the following assertions is true:
If there is a king in the hand then there is an ace.
If there is a king in the hand then there isn't an ace.

The model theory predicts that subjects using implicit models should respond that the king is more probable than the ace, which is the correct response. If the main connective is interpreted as a conjunction, however, the king is impossible because it yields a contradiction, and so subjects should respond that the ace is more likely. Only 10% of the subjects made this response.

Errors in reasoning in previous studies can be explained in terms of failures to retrieve appropriate rules of inference (e.g. Braine and O'Brien, 1991; Rips, 1994), or in terms of failures to consider all possible models of the premises (e.g. Johnson-Laird and Byrne, 1991). Illusory inferences, however, are not a result of such oversights. What is novel about them is that a conclusion that nearly everyone draws is totally wrong: what is judged more probable of two alternatives is impossible. The illusions are predicted by the model theory, but the illusory deductions appear to refute the theories based on rules of inference (Braine and O'Brien, 1991; Rips, 1994), which contain only rules that yield valid conclusions. Hence, these theories have no way to explain the systematically invalid conclusions that individuals draw to illusory inferences.

We have only just begun to explore the space of possible premises in search of illusory inferences. They are relatively rare, but there are illusions based on other connectives apart from exclusive disjunction, e.g. "and", and "if and only if". All the illusions seem to arise because human reasoners rely on implicit models, and so they overlook cases in which a state of affairs does not hold. To rely on as little explicit information as possible is a sensible solution to the all-pervasive problem of limited processing capacity. Just occasionally, however, the lack of explicit information leads human reasoners into the illusion that they grasp a set of possibilities that is in fact beyond them.

Acknowledgements

We thank Ruth Byrne, Jack Gelfand, Sam Glucksberg, Danny Kahneman, Geoffrey Keene, Joel Lachter, Rick Lewis, and Eldar Shafir, for their advice. The research was carried out with support from the James S. McDonnell Foundation and from Fonds pour la Formation de Chercheurs et l'aide a la Recherche (Quebec).

References

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

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

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

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