ELSEVIER

Acta Psychologica 93 (1996) 69–90

acta
psychologica

Illusory inferences about probabilities

P.N. Johnson-Laird *, Fabien Savary [1]

Department of Psychology, Princeton University, Princeton, NJ 08544, USA

Abstract

The mental model theory postulates that reasoners build models of the situations described in premises. A conclusion is *possible* if it holds in at least one model of the premises; it is *probable* if it holds in most of the models; and it is *necessary* if it holds in all of the models. The theory also postulates that reasoners represent as little information as possible in explicit models and, in particular, that they represent only information about what is true. One unexpected consequence of this assumption is that there should be a category of illusory inferences: they will have conclusions that seem obvious, but that are wholly erroneous. Experiment 1 established the existence of such illusory inferences about probabilities. Overall, 88% of the intelligent adult subjects chose as more probable an outcome that was impossible for at least one of the illusory problems. Experiment 2 corroborated the phenomenon and showed that illusory inferences include a wide variety of problems. Finally, the paper argues that current theories based on formal rules of inference are unlikely to be able to explain the illusions.

PsycINFO classification: 2340

Keywords: Reasoning; Mental models; Cognitive illusions; Formal rules of inference; Probability judgment

1. Introduction

How do people reason? They can undoubtedly do so – even without benefit of formal training – and yet they can say little about how their mental processes lead them to conclusions. Psychologists, however, have been less silent, and they have proposed a variety of theories (for a review, see Evans et al., 1993). The aim of the present paper is to advance our understanding of reasoning and, in particular, of reasoning about relative

* Corresponding author. E-mail: phil@clarity.princeton.edu, Tel.: +1 609 258-4432, Fax: +1 609 258-1113.

[1] E-mail: fabien@phoenix.princeton.edu.

PII S0001-6918(96)00022-4

probabilities. In order for readers to have an intuitive grasp of this sort of reasoning, we invite them to make the following two inferences, and to write down their answers for future reference.

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

> There is a king in the hand or there is an ace in the hand, or both.
> There is a queen in the hand or there is an ace in the hand, or both.
> Which is more likely to be in the hand: the king or the ace?

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

> If there is a jack in the hand, then there is a queen in the hand.
> If there is a ten in the hand, then there is a queen in the hand.
> Which is more likely to be in the hand: the queen or the jack?

We will return to these two problems later; meanwhile, readers should remember to write down their answers to them.

The principal division in psychological theories of reasoning echoes a distinction in logic. On the one hand, syntactic theories propose that reasoners rely on formal rules of inference akin to those of a logical calculus; on the other hand, semantic theories propose that reasoners construct model-like entities as their interpretations of premises. In previous publications, we have defended a psychological theory based on mental models and argued that it accounts for reasoning in all the main domains of deduction (see e.g. Johnson-Laird and Byrne, 1991). In the present paper, we describe how the two main sorts of psychological theories – formal rule theories and the mental model theory – can be extended to deal with reasoning about probabilities. Next, we report two experiments on reasoning about relative probabilities. Their results corroborate the model theory. The decisive evidence is that, as the theory predicts, there are illusory inferences about probabilities, i.e. there are certain premises from which most people draw the same conclusions, which seem obvious, and yet which are egregious errors.

2. Probabilities: Theories based on formal rules of inference

An inference is *valid* if its conclusion must be true given that its premises are true. Certain inferences are so obviously valid that many theorists have suggested that the mind is equipped with formal rules of inference that are used automatically to make these inferences. The paradigm case of such inferences is known as *modus ponens*, e.g.:

> If there is a queen in the hand then there is a jack in the hand.
> There is a queen in the hand.
> ∴ There is a jack in the hand.

Individuals draw this conclusion rapidly and without seeming to pause for thought. Rule theories accordingly postulate that the mind contains a formal rule of inference for modus ponens:

If p then q
p
$\therefore\ q$.

which is triggered by the premises and which leads at once to the derivation of the conclusion. Such theories were first proposed twenty years ago (see e.g. Osherson, 1975; Johnson-Laird, 1975; Braine, 1978), and they continue to flourish (see e.g. Braine and O'Brien, 1991; Rips, 1994). They predict that the difficulty of an inference depends on two main factors: the length of its formal derivation (using the rules postulated by the theory) and the ease of retrieving and using each of the required rules (which can be estimated from experimental data). These theories have had some success in fitting the results of experiments (Braine et al., 1984; Rips, 1983, 1994).

Formal rule theories have so far been formulated to deal only with deductions, such as modus ponens, which lead to conclusions that are necessarily true. In contrast, the following problem concerns the relative probabilities of two events:

If there is a jack or queen in the hand, then there is an ace.
Which is more likely to be in the hand: the jack or the ace?

It has a valid answer, that is, an answer that must be true given that the premise is true: the ace is more likely to be in the hand than the jack. To the best of our knowledge, psychologists have not previously studied such problems, and it is not clear whether formal rule theories are intended to apply to them. Yet, they could be made to yield judgments of relative probability in the following way. Given the premise of the problem above:

If there is a jack or queen in the hand, then there is an ace

reasoners could proceed in the following way. They start by making a supposition, i.e. assumption for the sake or argument:

There is a jack in the hand.

Next, according to certain formal rule theories (e.g. Braine and O'Brien, 1991), they could derive the conclusion:

There is an ace in the hand

using a rule of the form:

p
if p or q then r
$\therefore\ r$

Similarly, from the supposition:

There is a queen in the hand

they can derive the same conclusion using a variant of the same rule. A system of 'bookkeeping' could keep track of the respective possibilities, i.e. whenever there is a jack, there is an ace, but not vice versa, and hence yield the conclusion: an ace is more likely to be in the hand than a jack. No such theory has yet been proposed by any formal theorist, and so we will not pursue the details any further. For our purposes, it is sufficient to know that a rule theory for drawing conclusions about relative probabilities is at least feasible.

3. Probabilities: Theories based on mental models

According to the mental model theory, reasoning is a semantic process rather than a syntactic one. Reasoners imagine the states of affairs that satisfy the premises, that is, they build mental models of the relevant situations based on their understanding of the premises and on general knowledge; they formulate a conclusion that is true in these models; and they establish its validity by ensuring that there are no models of the premises in which the conclusion is false (Johnson-Laird, 1983). By a mental model, we mean a representation that corresponds to a set of situations, and that has a structure and content that captures what is common to these situations. The contrast between formal, syntactic, methods in logic and semantic methods is a familiar one; and logicians have shown that there exists a sharp division between 'proof theoretic' methods based on formal rules and 'model theoretic' methods based on semantics. In certain branches of logic, proof theory is *incomplete* in that one cannot formulate a consistent set of rules that captures all and only the valid inferences (see e.g. Jeffrey, 1981).

One advantage of the mental model theory is that it provides a unified account, so far lacking in formal rule theories, of logical reasoning that leads to necessary conclusions, probable conclusions, and possible conclusions. A conclusion is necessary – it must be true – if it holds in all the models of the premises; a conclusion is probable – it is likely to be true – if it holds in most of the models of the premises; and a conclusion is possible – it may be true – if it holds in at least some model of the premises (Johnson-Laird, 1994). The theory purports to explain how intelligent, but mathematically ignorant, individuals – such as Aristotle! – reason about probabilities. Certain numerical judgments of probability may be based on the frequency of models in which events occur – a point to which we return later; other judgments may be based on the availability of models, i.e. how easy it is to construct them (see Tversky and Kahneman, 1973); and still others may call for models to be linked to numerical representations of probabilities. It remains to be seen how far the theory can be extended to cope with numerical probabilities.

The fundamental *representational* assumption of the mental model theory is that individuals seek to minimize the load on working memory by representing explicitly only those cases that are true. Thus, a simple conjunction:

There is a king in the hand and there is a ace 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
```

There is no need to represent explicitly cases where the conjunction is false. 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:

$$\begin{array}{ll} k & \\ & a \end{array}$$

where each line represents a separate model. In this case, even those components of the assertion that would be false in these models are not explicitly represented, that is, the models do not explicitly represent 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:

$$\begin{array}{ll} k & \neg a \\ \neg k & a \end{array}$$

where '$\neg$' denotes negation. It is these negative elements, $\neg a$ and $\neg k$, that people do not ordinarily represent explicitly.

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 the antecedent is false, i.e. where there is *not* a king in the hand, which they represent in a wholly implicit model denoted here by an ellipsis:

$$\begin{array}{ll} k & a \\ \ldots & \end{array}$$

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. The representation of a biconditional:

There is a king if, and only if, there is an ace

has exactly the same initial models, but reasoners need to make a mental footnote that both the king and the ace are exhaustively represented in the explicit model.

4. Models and illusory inferences about probabilities

The model theory predicts that conclusions which hold in most of the models of the premises will be judged to be probable. It also makes predictions about judgments that one event is more probable than another. There are two potentially relevant principles.

> Principle 1: if A occurs in each model in which B occurs, but B does not occur in each model in which A occurs, then A is more likely than B. In other words, if the models in which A occurs contain the models in which B occurs as a proper subset, then A is more likely than B.

Consider again the example:

> If there is a jack or a queen in the hand, then there is an ace.
> Which is more likely to be in the hand: the jack or the ace?

The antecedent of the conditional:

> There is a jack or a queen in the hand

calls for the models:

```
j
     q
j    q
```

given an inclusive interpretation of the disjunction. These models are embedded in the interpretation of the conditional:

```
j         a
     q    a
j    q    a
    ...
```

The models in which the ace occurs contain as a proper subset the models in which the jack occurs, and so the ace is more likely to be in the hand than the jack. Principle 1 is, of course, valid provided that one takes into account all of the possible models of the premise. The second principle is simpler and more general, and it includes principle 1 as a special case:

> Principle 2: if A occurs in more models than B, then A is more probable than B.

This principle is risky. It is valid only if each model is equally probable, i.e., each model corresponds to a set of situations, and each of these sets is equally probable.

The model theory makes a striking prediction. There are premises with initial models that support grossly erroneous conclusions. Hence, these premises should give rise to illusory inferences, i.e. nearly everyone should draw the same conclusion, it should seem obvious, and yet it is completely wrong. Readers have already encountered two illusory inferences (Problems 1 and 2 in the Introduction). Readers may have responded to Problem 1 that the ace is more likely to be in the hand than the king. Likewise, they may have responded to Problem 2 that the queen is more likely to be in the hand than the jack. In either case, they have succumbed to an illusion. It is impossible for an ace to be in the hand in the first problem, and it is impossible for a queen to be in the hand in the second problem. We will first outline the model theory's predictions about these two inferences, and then explain the correct conclusions.

Consider problem 1, which we abbreviate as:

Only one is true:
King or ace, or both.
Queen or ace, or both.
Which is more likely: king or 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 means that either one assertion or the other assertion is true, but not both of them. That is, it calls for an exclusive disjunction of them, and the models for an exclusive disjunction, X or else Y, are:

X	
	Y

and so the disjunction calls for a list of all the models in the two alternatives. Hence, the problem as a whole calls for the following models:

k		
		a
k		a
	q	
		a
	q	a

If subjects estimate probabilities using the second riskier principle of the two described above, then they judge the probability of an event on the basis of the proportion of models in which it holds. They will therefore respond that the ace is more probable than the king. If, however, subjects assume that the models may not be equiprobable, they will conclude that the problem is indeterminate, e.g. the probability of the king alone could be greater than the probabilities of all the other models summed together. Both of these responses are wrong, however.

What has gone wrong? If only one of the two assertions is true, then the other assertion is false: the two premises are in an exclusive disjunction, and so when one is true, the other is false. The models, however, represent only the true cases. When the false cases are taken into account, the correct answer emerges. 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*. Either way, there is no ace – it cannot occur in the hand. Hence, the king, which can occur in the hand, is more probable than the ace, which cannot occur in the hand.

Now, consider problem 2, which we abbreviate.

Only one is true:
If jack then queen.
If ten then queen.
Which is more likely: queen or jack.

Once again, if readers answered that the queen is more likely than the jack, then they succumbed to an illusion. It arises because an exclusive disjunction calls for listing the two sets of models:

```
j          q
      10   q
      ...
```

and now even the sound first principle for estimating probabilities dictates that the queen is more probable than the jack. But, as with the first problem, the correct answer depends on bearing in mind the false contingencies, that is, when one conditional is true the other is false. The first conditional is false when there is a jack but not a queen, and the second conditional is false when there is a ten but not a queen. Either way, there is not a queen: it is impossible, but the jack is not impossible, and so the correct answer is that the jack is more probable than the queen.

5. Experiment 1

No previous experiments, as far as we know, have examined logical inferences about relative probabilities. The first aim of Experiment 1 was accordingly to determine whether logically-untrained individuals could make such inferences. The second aim was to test the model theory's prediction that certain of these inferences are illusory and that the subjects would therefore be prone to err on them. The experiment compared the two illusions described above with two simpler control problems, which should not be illusory according to the model theory. The third aim was to assess whether subjects based their judgments of relative probability on principle 1 concerning proper subsets or on the more general, but riskier, principle 2 based on the assumption of equiprobability.

5.1. Method

5.1.1. Design and materials

The subjects acted as their own controls and carried out two illusory inferences stated here in abbreviated form:

1. Only one assertion is true about a specific hand of cards:
   King or ace, or both.
   Queen or ace, or both.
   Which is more likely: king or ace?
2. Only one assertion is true about a specific hand of cards:
   If king then ace.

If queen then ace.
Which is more likely: ace or king?

In fact, each problem concerned different cards, but for convenience we have stated the problems as though they were always about the same particular cards. The subjects also carried out two control problems:

3. If king then ace.
   Which is more likely: king or ace?
4. If king or queen then ace.
   Which is more likely: ace or king?

Control problem 3 calls for the initial models:

```
k       a
...
```

with a footnote to the effect that kings are exhausted in the explicit model, i.e. kings cannot occur in any other model, whereas aces can occur in the model signified by the ellipsis. Subjects should therefore infer that the ace is more likely to be in the hand than the king. If subjects make a biconditional interpretation, however, then they will treat the initial model as exhaustively representing both kings and aces, and so they will judge the two cards to be equiprobable.

Control problem 4 calls for the initial models:

```
k           a
      q     a
k     q     a
      ...
```

The third of these models will be omitted by those individuals who interpret 'or' as an exclusive disjunction. But, in either case, the models containing kings are a proper subset of the models containing aces, and so the subjects should judge that the ace is more likely to be in the hand than the king. These conclusions to the two control problems are correct, that is, even when the models are made completely explicit they still support the same conclusions.

Each subject carried out the four inferences in a different order, i.e. one of the 24 possible orders. The materials for each problem concerned a specific hand of cards, and four distinct hands of cards were assigned to each problem at random.

5.1.2. Procedure

The subjects were tested individually. They were told that their task was to make a series of judgments about how likely one card or another was to be in a hand of cards. They were to base their judgments solely on the information given to them. They could take as much time as they needed to make their response. The typed problems were presented on separate pages of paper, and the subjects wrote their answers beneath the problems.

5.1.3. Subjects

Twenty-four Princeton students carried out the experiment. None of them had received any training in logic or had participated in any experiments on reasoning. They

Table 1
The percentages of responses to illusory and control inferences in Experiment 1. The correct responses are shown in bold print

Type of problem		Responses and their percentages		
		ace	king	equiprobable
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

were paid $4 per hour for participating in the experiment, which lasted for about ten minutes.

5.2. Results

The results are summarized in Table 1. There was a massive difference between the two sorts of inferences: the subjects were correct on 71% of the control inferences, but on only 17% of the illusory inferences. 20 out of the 24 subjects were correct on more of the control inferences than the illusory ones, and there were two ties (Sign test, $p < 0.001$). 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.

5.3. Discussion

The experiment confirmed the existence of illusory inferences and it also established that individuals judge relative probabilities according to a risky principle (principle 2 above): they judge that one event is more probable than another if it occurs in more more models than the other event. The risk here is that one model may be much more probable than another, and so the procedure is sound only if the principle of 'indifference' is correct, i.e. there is no reason to suppose that one model is more probable than another.

The model theory rests on the assumption that individuals focus on true states of affairs, which they represent explicitly, and rapidly forget, if they represent them at all, false states of affairs. This assumption led to the prediction of the existence of illusory inferences, but perhaps our results are merely a happy coincidence, and the true cause of the illusions is quite different from model theory's account. We will discuss the possibility of such alternative explanations later. However, we designed Experiment 2 in part to examine their plausibility.

6. Experiment 2

Experiment 2 was designed to examine a variety of possible illusions, including those that depend on two main sorts of major connective: biconditionals and exclusive disjunctions, and those that depend on a variety of minor connectives, including conditionals, biconditionals, and exclusive disjunctions. We were particularly interested in finding minimal illusions, that is, those based on the simplest possible premises. Hence, four of the illusory inferences were based on only two connectives, while the remaining two illusory inferences were based on three connectives. Unlike the previous experiment, the matched control problems in all cases had the same major connectives as the illusory problems, and, as far as possible, the same minor connectives. The procedure was also more sensitive, because for each problem the subjects made independent estimates of the probabilities of the two cards (making their responses by clicking a mouse to mark their estimates on separate scales presented on a computer screen). In a separate part of the experiment, the subjects were asked to consider each problem and to state what conclusion, if any, followed from the premises. The results of this part of the experiment, however, were not sufficiently revealing to merit a full discussion, and so we deal with them only briefly.

6.1. Method

6.1.1. Design

The subjects inferred probabilities for six illusory problems and six matched control problems. In order to make the problems easy to understand, there were two separate assertions in each problem, and the main connective in a problem was expressed in the following way:

Exclusive disjunction: ''Only one of the following assertions is true about a specific hand of cards''.

Biconditional: ''If one of the following assertions is true about a specific hand of cards, then so is the other assertion''.

Two of the matched pairs of problems depended on three sentential connectives, and four of the matched pairs of problems depended on two sentential connectives.

The illusory problem in the first pair of problems based on three connectives had an exclusive disjunction as its main connective:

1. *Only one* of the following assertions is true about a specific hand of cards:
   If there is a jack in the hand then there is a queen in the hand.
   If there isn't a jack in the hand then there is a queen in the hand.

The subjects' task was to estimate the probability that the jack was in the hand and to estimate separately the probability that the queen was in the hand. The subjects made both responses by indicating the relevant position on two separate computer-presented scales running from 'impossible' to 'certain'. According to the model theory, this problem should elicit the initial models:

$$\begin{array}{ll} j & q \\ \neg j & q \\ \ldots & \end{array}$$

and so subjects should assign a higher probability to the queen than to the jack. It is by no means certain whether individuals will represent the implicit model signified here by the ellipsis. The implicit model may be forgotten, or it may be omitted because the two antecedents of the conditionals exhaust the possibilities. In either case, the models support the same inference: the queen should be assigned a higher probability than the jack. This conclusion, of course, is an illusion: one of the premises must be false, and so there cannot be an queen. The fully explicit models for the premise are as follows:

$$\begin{array}{ll} \neg j & \neg q \\ j & \neg q \end{array}$$

The correct response is accordingly to assign the jack a higher probability than the queen, whose presence in the hand is impossible.

The matched control problem took the form:

1′. *Only one* of the following assertions is true about a specific hand of cards:
If there is a jack in the hand then then there is a queen in the hand.
If there is a jack in the hand then there is not a queen in the hand.

We abbreviate the statement of this problem as follows:

Only one is true:
If J then Q.
If J then not Q.

using the conventions that 'J' denotes 'there is a jack', and 'not Q' denotes 'there is not a queen'. This problem should elicit the following models:

$$\begin{array}{ll} j & q \\ j & \neg q \\ \ldots & \end{array}$$

and so subjects should assign a higher probability to the jack than to the queen. The fully explicit models are in this case:

$$\begin{array}{ll} j & q \\ j & \neg q \end{array}$$

Hence, the response is correct.

The second pair of problems with three connectives were based on biconditionals. According to the model theory, both problems yield only implicit models, that is, models with no explicit content (see problems 2 and 2′ in Table 2), and so they were included in the experiment primarily to see how subjects would respond when a problem seemed not to have any explicit model. In the illusory case, the correct response is that the queen is more probable than the jack; in the control case, the correct response is unclear because the premises are self-contradictory.

We also used four illusory problems and four matched controls based on only two sentential connectives. For two of these pairs, the main connective was an exclusive disjunction, and for the other two of these pairs, it was a biconditional. One of the illusory problems based on a biconditional was:

3. If one of the following assertions is true about a specific hand of cards then so is the other assertion:
There is a jack if and only if there is a queen.
There is a jack.

We can again abbreviate this problem:

If one is true so is other:
J iff Q.
J.

It should elicit the models:

$$\begin{array}{ll} j & q \\ \ldots & \end{array}$$

and so subjects should estimate that the two cards have equal probabilities of occurring in the hand. But, the fully explicit models of the problem are:

$$\begin{array}{ll} j & q \\ \neg j & q \end{array}$$

and so the correct answer is that the queen is certain to be in the hand whereas the jack is not.

The matching control problem was of the form:

3′. If one is true so is other:
If J then Q.
J.

which should elicit the single model:

$$\begin{array}{ll} j & q \end{array}$$

and so subjects should again infer that the two cards have equal probabilities of occurring in the hand. In this case, the models are correct, and so the conclusion is, too. The full set of six illusory problems and their matched controls are shown in Table 2. The problems were presented in a different random order to each subject.

6.1.2. Procedure and materials

The subjects were tested individually in a quiet room. The experimenter explained that the purpose of the experiment was to elicit judgments about which cards were more likely to be in a specific hand of cards, and that these estimates were to be based solely

Table 2
The six illusory problems and their matched controls in Experiment 2. The models that subjects should construct are on the left, and the fully explicit correct models are on the right. 'Iff' denotes 'if and only if'

	Illusory			Control	
1.	Only one is true: If J then Q. If not J then Q.		1′.	Only one is true: If J then Q. If J then not Q.	
	j q	¬j ¬q		j q	j q
	¬j q	j ¬q		j ¬q	j ¬q
	...			...	
2.	If one is true so is other: J and Q. J and not Q.		2′.	If one is true so is other: J iff Q. J iff not Q.	
	...	¬j ¬q		...	null (i.e. one premise
		¬j q			contradicts the other.)
3.	If one is true so is other: J iff Q. J.		3′.	If one is true so is other: If J then Q. J.	
	j q	j q		j q	j q
	...	¬j q			
4.	If one is true so is other: J or else not Q. J.		4′.	If one is true so is other: J or else Q. Not Q.	
	j	j q		j ¬q	j ¬q
	j ¬q	¬j q		...	j q
	...				
5.	Only one is true: Not J or else not Q. Q.		5′.	Only one is true: Not J or else Q. Not J.	
	¬j	j ¬q		¬j	j q
	¬q	j q		q	¬j q
	q				
6.	Only one is true: J iff not Q. J.		6′.	Only one is true: not J iff not Q. J.	
	j ¬q	¬j q		¬j ¬q	¬j ¬q
	j	j q		j	j ¬q

on the information presented in the verbal statement of the problems. He then showed the subjects how to use the micro-computer and the mouse (an Amiga 2000 computer running a program written in Basic to control the experiment). The subjects carried out one practice trial in which they made two simple assessments of probability. Each trial began with the verbal presentation of the premises with two horizontal scales presented below – one scale for one card, and the other scale for the other card. The scales were labeled at equal intervals with the following terms:

impossible unlikely 50/50 likely certain

The subjects made their responses by moving a mouse that in turn directed the cursor to the desired point on the scale. The subject then clicked on the mouse, and the scale was divided with a vertical line at that point. The program recorded the distance of the subject's response along the scale. After the subject had made both responses to a problem, the trial ended, and the next trial began when the subject clicked in the ready box. The particular cards for each problem were assigned at random, but no problem had the same pair of cards as any other.

In the other part of the experiment, the subjects were asked to state what followed from each of the problems, that is, they had to write down what conclusion, if any, *must* be true given the information in the premises. This part of the experiment was administered as a simple paper-and-pencil test, and the twelve problems were presented in a random order.

6.1.3. Subjects

Twenty Princeton students took part in the experiment. None of them had any training in logic or had participated in any previous experiments on reasoning. They were paid $5 per hour to carry out the experiment, which lasted for approximately 40 minutes.

6.2. Results

For our first analysis of the probability judgments, we divided them into three categories:

1. jack assigned a higher probability than queen (J > Q)
2. queen assigned a higher probability than jack (Q > J)
3. jack and queen assigned equal probabilities (J = Q)

where, for convenience, we have assumed that each problem concerned a jack and a queen. We then scored the number of correct responses for each of the problems. Control problem 2′ has no clear correct answer because it expresses a contradiction, and so we dropped both it and its matching illusory problem 2 from this analysis. Table 3 presents the percentages of correct responses for each of the remaining ten problems. Overall, the subjects made 13% correct responses to the illusory problems, but 64% correct responses to the control problems. The difference between the two conditions was highly reliable: 17 subjects made fewer correct inferences to the illusory problems than to the control problems, 1 subject had the opposite pattern of results – this subject made only one correct response, and there were two ties (Wilcoxon's $T = 1$, $n = 18$, $p \ll 0.005$). Likewise, all five pairs of materials showed the predicted difference (Sign test, $p < 0.04$).

The subjects' responses matched the model theory's predictions on 56% of the illusory problems and, of course, on 64% of the control problems. In the case of problems that are not illusory, the model theory predicts that the greater the number of explicit models that have to be constructed in order to respond correctly, the harder the task should be (Johnson-Laird and Byrne, 1991). This prediction is corroborated by the control problems: problems 3′ and 4′ require only one explicit model whereas problems 1′, 5′, and 6′ require two explicit models (see Table 2). The former elicited 83% correct

Table 3
The correct responses and their percentages for five illusory problems and their matched controls in Experiment 2. J > Q indicates that the jack was given a higher probability than the queen, Q > J indicates that the queen was given a higher probability than the jack, J = Q indicates that the jack and the queen were given the same probabilities. Problems 2 and 2′ were not included in this analysis (see text)

	Illusory			Control	
1.	Only one is true: If J then Q. If not J then Q. J > Q:	15	1′.	Only one is true: If J then Q. If J then not Q. J > Q:	55
3.	If one is true so is other: J iff Q. J. Q > J:	10	3′.	If one is true so is other: If J then Q. J. J = Q:	95
4.	If one is true so is other: J or else not Q. J. Q > J:	5	4′.	If one is true so is other: J or else Q. Not Q. J > Q:	70
5.	Only one is true: Not J or else not Q. Q. J > Q:	20	5′.	Only one is true: Not J or else Q. Not J. Q > J:	35
6.	Only one is true: J iff not Q. J. Q > J:	15	6′.	Only one is true: not J iff not Q. J. J > Q:	65
	Overall percentages:	13			64

responses, whereas the latter elicited only 52% correct responses, and the difference was reliable (Wilcoxon's $T = 9$, $n = 16$, $p \ll 0.005$).

Each problem called for a subject to infer the relative probabilities of two cards, and 89% of these judgments aligned a card on one of the five canonical points labeled on the scale (allowing for an error of no more than two hundredths of the scale). Four out of the twenty subjects were responsible for just over two thirds of the judgments that were not on one of the five canonical points, and nine subjects made only canonical judgments. The percentages of judgments were as follows:

Impossible:	16
unlikely:	3
50/50:	33
likely:	3
certain:	34

We have pooled the data for the illusory and the control problems because their patterns were highly similar. As the percentages show, the subjects overwhelmingly judged the cards as impossible, 50/50, or certain. This distribution of responses is to be expected from the models of the premises: cards in the models are certain, or impossible, or occur in one of two models but not the other (see Table 2).

In the other part of the experiment, the subjects were asked to state what conclusion, if any, followed from each problem. The results were relatively noisy, perhaps because the subjects were reluctant to draw conclusions corresponding to a categorical premise (even though it occurred as part of an exclusive disjunction or biconditional). They were revealing, however, about problems 2 and 2′, where the model theory predicts that subjects construct only an implicit model in both cases (see Table 2). Subjects had a tendency to conclude that there was a contradiction (60% of responses were of this form for problem 2 and 50% of responses were of this form for problem 2′). In fact, the illusory problem is not self-contradictory, but the control problem is self-contradictory.

6.3. Discussion

The results showed that the phenomenon of illusory inferences occurs with a variety of connectives. The main connective can be an exclusive disjunction or a biconditional. For example, given problem 1:

1. Only one is true:
   If J then Q.
   If not J then Q.

few subjects grasped that the queen was impossible, and the modal response was that the queen was more probable than the jack. The experiment also showed that illusions could be created in a minimal way by assertions containing only two connectives, e.g.:

3. If one is true so is other:
   J iff Q.
   J.

where most subjects erroneously inferred that the two cards had the same probabilities of being in the hand, though in fact the queen is certain to be in the hand but the jack is not.

According to the model theory, the illusions arise because reasoners represent true cases, but not false cases. The difference between the illusory and the control problems is simply that this tendency has no effect on correctness in the case of the controls. The control problems were selected to be as similar as possible to the illusory problems in the choice of connectives and negatives. One consequence was that some of the control problems called for two distinct models to be constructed in order to reach the correct answer, whereas others called for only one model. Previous studies of reasoning with sentential connectives have shown that two model problems can be difficult for subjects (see Johnson-Laird et al., 1992). Our results with the control problems corroborated this phenomenon: the two-model control problems were reliably harder than the one-model control problems.

7. General discussion

The model theory was originally developed as an account of how people draw logically-necessary conclusions. However, it also gives a simple explanation of how reasoners reach conclusions about what is probable (Johnson-Laird, 1994): a situation is probable if it holds in most models of the premises, and, assuming a principle of 'indifference' according to which models are equiprobable, one event is more probable than another if it occurs in more models than the other. The twist in these predictions is that reasoners are likely to construct explicit models only of true cases and to forget about the false cases, especially if the premises are complex. Where these models yield a different conclusion from the one supported by fully explicit models of both true and false components, subjects should draw illusory inferences: their conclusions should be totally wrong. It is worth emphasizing that the prediction is based on the models needed for deduction, that is, we did not develop a new theory of models to account for reasoning about probabilities.

Experiment 1 confirmed the existence of illusory inferences. It also showed that subjects do tacitly assume a principle of 'indifference' and infer that whichever event occurs in more models is the one that is more likely to occur. Thus, given problem 1:

Only one assertion is true:
  king or ace, or both.
  queen or ace, or both.
Which is more likely: king or ace?

Most subjects inferred that the ace is more likely. This illusory inference rests on the construction of the following set of models:

$$\begin{array}{ccc} k & & \\ & & a \\ k & & a \\ & q & \\ & & a \\ & q & a \end{array}$$

in which there are more models containing the ace than models containing the king. The fully explicit models for this problem, however, are as follows:

$$\begin{array}{ccc} \neg k & q & \neg a \\ k & \neg q & \neg a \end{array}$$

The ace is impossible and so less likely to occur in the hand than the king.

Experiment 2 provided further corroboration of the model theory's predictions by showing that illusory inferences occur with a variety of different sorts of connectives. It also established that illusions can be constructed with just two sentential connectives. The superficial similarity between these illusory problems and their matched control problems is quite striking. For example, a cursory examination of the following pair of problems:

4. If one is true so is other:	4′. If one is true so is other:
J or else not Q.	J or else Q.
J.	Not Q.

is unlikely to suggest that reasoners will perform in a radically different way with them. Yet, only one subject made the correct judgments about problem 4, i.e. the queen has a higher probability of being in the hand than the jack, whereas 14 subjects made the correct judgments for problem 4′, i.e. the jack has a higher probability of being in the hand than the queen.

Is there any obvious alternative explanation for subjects' susceptibility to the illusions? Colleagues who have succumbed to the illusions – and they include a number of distinguished cognitive psychologists – have suggested several alternative explanations for them. We will consider three possibilities. First, subjects may misinterpret an assertion of the form:

Only one of the following assertions is true

to mean:

One of the assertions is true and the other is of an unknown truth value

and then reason in a wholly correct way. Likewise, they may misinterpret an assertion of the form:

If one of the following assertions is true about a specific hand of cards, then so is the other assertion

to mean:

Either both of the assertions are true or else they have unknown truth values.

Given a disjunction of two assertions, X and Y, the first of these hypotheses implies that individuals consider one case in which X is true and Y is either true or false, and another case in which Y is true and X is either true or false. When these cases are spelt out explicitly, they are as follows:

X	Y
X	¬Y
¬X	Y

In other words, the interpretation is equivalent to an *inclusive* disjunction of the two assertions, X and Y. However, an inclusive disjunction of the two conditionals in problem 1 of Experiment 2:

If J then Q.
If not J then Q.

yields a tautology: there is either a jack or not a jack, and there is either a queen or not a queen. Hence, there is no reason to infer that the queen is more likely than the jack – yet subjects made exactly that inference, and so it follows they are not treating the two conditionals as being in an inclusive disjunction.

Given two assumptions, X and Y, the second of these two hypotheses implies that subjects consider one case in which they are both true, and other cases in which they are each either true or false. This treatment yields a tautology based on X and Y:

X	Y
X	¬Y
¬X	Y
¬X	¬Y

Once again, however, this interpretation provides no basis for inferring that one card is more probable than another, and so it is refuted by the results of Experiment 2 (see problem 6′).

Second, subjects may have a formal rule of inference that converts an exclusive disjunction of two conditionals:

> If there is a king then there is an ace.
> If there isn't a king then there is an ace.

into a single conditional with a disjunctive antecedent:

> If there is a king or isn't a king then there is an ace.

For problems of this sort, the hypothesis makes the same prediction as the model theory, because it postulates that subjects construct the models:

k	a
¬k	a

The difficulty for the formal rule, however, is that it cannot explain the other sorts of illusion, such as the ones where the main connective is a biconditional.

Third, subjects may have interpreted the two premises in an illusory inference as though they were in a conjunction, and then reasoned in a wholly correct way. In our view, it is again unlikely that intelligent students would take the assertion equivalent to an exclusive disjunction:

> *Only one* of the following assertions is true

to mean:

> *Both* of the following assertions are true.

But, it is feasible that the subjects took the biconditional assertion:

> If one of the following assertions is true about a specific hand of cards then so is the other assertion

to mean that both assertions are, in fact, true. Experiment 2 refutes the first of these hypotheses. The control problem 1′:

> Only one of the following assertions is true:
> If jack then queen.
> If jack then not queen.

would yield the inference that the jack is impossible, because its presence would yield a contradiction. Only 10% of subjects judged that the queen had a higher probability than the jack, whereas 55% of subjects made the response predicted by the model theory: the jack is more probable than the queen (see Table 3). The treatment of a biconditional as though it were a conjunction is similar to the model theory's account, which postulates that reasoners represent explicitly the case where the two assertions are true, and that they represent the case where they are both false with only an implicit model (signified by the ellipsis in our notation). Indeed, if subjects omit the implicit model, or forget it, then the two accounts are one and the same. The implicit model, however, does seem to be necessary in order to explain how subjects make the following sort of inference based on the more conventional expression of a biconditional:

There is a king if, and only if, there is a queen.
There isn't a queen.
∴ There isn't a king.

Many reasoners are able to make this inference, presumably by fleshing out the content of the implicit model explicitly before they take into account the information in the second premise (see Johnson-Laird et al., 1992).

In general, the alternative explanations can account for only *some* of our data, whereas the model theory predicts the existence of the illusory inferences in general. The illusions refute the extension of formal rule theories to deal with probabilities – at least, the extension that we described earlier, because it yields only valid inferences. Of course, one could invoke a different (invalid) formal rule to deal with each of the different sorts of illusory inference, but such an account would be entirely *post hoc*. It might well lead to invalid inferences that reasoners do not, in fact, make. Moreover, there is a general problem in invoking invalid rules to explain the illusions. If human reasoners were guided by such a system, they would be intrinsically irrational and their capacity for rational thinking as manifest in mathematics and logic would be wholly inexplicable. In contrast, the model theory assumes that human reasoners are rational in principle because they grasp that an argument is valid if, and only if, there are no counterexamples to it, i.e. no models of the premises in which the conclusion is false. They err in practice, however, because their working memory is limited and they tend to represent explicitly only true cases.

Most errors in reasoning can be explained in terms of failures to use 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). The illusory inferences in the present experiments, however, are not a result of such oversights. They yield conclusions that nearly everyone draws, yet that are totally wrong, e.g. what is judged more probable of two alternatives is impossible. Errors in Wason's 'THOG' task have a similar pattern (Wason, 1977). We have also shown in an unpublished study that illusions occur in deductions yielding necessary conclusions. This phenomenon is contrary to current theories based on rules of inference (e.g. Braine and O'Brien, 1991; Rips, 1994), just as our present results refute an extension of formal rules to deal with probable conclusions. Current theories use only rules that yield valid conclusions, and so they have no way to explain the systematically invalid conclusions that individuals drew to illusory inferences. Rule theorists could well follow Jackendoff (1988) and invoke unsound rules that deliver invalid conclusions. Rips (1994) clearly countenances the possibility: "If people possess ... normatively inappropriate rules for reasoning with uncertainty, it seems a short step to assuming that they have similarly inappropriate rules for reasoning deductively" (p. 383). It remains to be seen whether anyone will succeed in formulating a rule theory that falls into deductive illusions but copes satisfactorily with the control problems.

We have just begun to explore illusory inferences and remedial procedures that render people less susceptible to them. Illusions are relatively rare in the 'space' of possible inferences, but there are many sorts of them. If the model theory is on the right lines, they arise because reasoners overlook cases in which a state of affairs is false. 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, it leads us into a profound illusion about what is probable.

Acknowledgements

We thank Ruth Byrne, Jonathan Evans, Jack Gelfand, Vittorio Girotto, Sam Glucksberg, Tony Greenwald, Geoff Hinton, Danny Kahneman, Geoffrey Keene, Joel Lachter, Paolo Legrenzi, Alan Leslie, Rick Lewis, Bonnie Meyer, Eldar Shafir, Patrizia Tabossi, and Amos Tversky, for their helpful suggestions. 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). We are grateful to the participants of SPUDM-15 (Subjective Probability, Utility, and Decision Making) for their comments. Experiment 1 was presented to the 17th Annual Conference of the Cognitive Science Society, Pittsburgh, July, 1995.

References

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

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

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

Evans, J.St.B.T., S.E. Newstead and R.M.J. Byrne, 1993. Human reasoning: The psychology of deduction. Hillsdale, NJ: Erlbaum.

Jackendoff, R., 1988. 'Exploring the form of information in the dynamic unconscious'. In: M.J. Horowitz (Ed.), Psychodynamics and cognition. Chicago, IL: University of Chicago Press.

Jeffrey, R., 1981. Formal logic: Its scope and limits (2nd ed.). New York: McGraw-Hill.

Johnson-Laird, P.N., 1975. 'Models of deduction'. In: R.J. Falmagne (Ed.), Reasoning: Representation and process in children and adults. Hillsdale, NJ: Erlbaum.

Johnson-Laird, P.N., 1983. Mental Models: Towards a cognitive science of language, inference and consciousness. Cambridge: Cambridge University Press/Cambridge, MA: Harvard University Press.

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

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

Johnson-Laird, P.N., R.M.J. Byrne and W.S. Schaeken, 1992. Propositional reasoning by model. Psychological Review 99, 418–439.

Osherson, D., 1975. 'Logic and models of logical thinking'. In: R.J. Falmagne (Ed.), Reasoning: Representation and process in children and adults. Hillsdale, NJ: Erlbaum.

Rips, L.J., 1983. Cognitive processes in propositional reasoning. Psychological Review 90, 38–71.

Rips, L.J., 1994. The psychology of proof. Cambridge, MA: MIT Press.

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

Wason, P.C., 1977. 'Self-contradictions'. In: P.N. Johnson-Laird and P.C. Wason (Eds.), Thinking: Readings in cognitive science. Cambridge: Cambridge University Press.