COGNITION

Cognition 71 (1999) 191–229

# Illusory inferences: a novel class of erroneous deductions

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

[a]*Department of Psychology, Princeton University, Princeton, NJ 08544-1010, USA*
[b]*3040 Henri de Salieres #3, Montréal, Québec, Canada H1N 2Y2*

Received 26 February 1998; accepted 16 February 1999

**Abstract**

The mental model theory postulates that reasoners build models of the situations described in premises, and that these models normally make explicit only what is true. The theory has an unexpected consequence: it predicts the occurrence of inferences that are compelling but invalid. They should arise from reasoners failing to take into account what is false. Three experiments corroborated the systematic occurrence of these illusory inferences, and eliminated a number of alternative explanations for them. Their results illuminate the controversy among various current theories of reasoning. © 1999 Elsevier Science B.V. All rights reserved.

*Keywords:* Reasoning; Mental model theory; Illusory inferences

## 1. Introduction

> Science can never grapple with the irrational.
>
> (Oscar Wilde: *An Ideal Husband*)

The ability to make deductions is a central component of human thinking (Rips, 1994). Without it, there would be no science, mathematics, logic, laws, or other

* Corresponding author. Tel.: +1-609-2584432; fax: +1-609-2581113;
*E-mail address*: phil@clarity.princeton.edu (P.N. Johnson-Laird)

PII: S0010-0277(99)00015-3

general principles. Human beings are able to reason, however, even if they have had no training in logic. What is problematic is the nature of the mental processes that underlie their reasoning. When individuals ‘think aloud’ as they reason, their protocols allow one to infer their high-level strategies but not their underlying inferential processes. They say, for instance, that they are considering a particular premise, or that they are making an assumption, but they say nothing about *how* they derive a particular conclusion. This introspective void has been filled by a variety of psychological theories of deductive reasoning, and the controversy amongst them is hot (for a review, see Evans et al., 1993). Some of these theories rely on a memory for specific examples or for rules with a content tailored to particular domains of expertise. Yet, human reasoners can make deductions about matters for which they have no specific knowledge, and the controversy is hottest about this general logical ability.

Logicians have devised various logical calculi to help them to assess the validity of inferences. The so-called ‘sentential’ calculus, for example, concerns the logic of negation and idealized sentential connectives, such as ‘if’, ‘or’, and ‘and’. Arguments in the sentential calculus can be evaluated in at least two dis-tinct ways. The first way is known as ‘proof-theoretic’ (see e.g. Jeffrey, 1981). The calculus is laid down using formal rules of inference, and arguments are derived in formal proofs that start with the premises and lead to the conclusion in a series of steps, each sanctioned by a rule of inference. The paradigm of a formal rule of inference is *modus ponens*, which allows inferences of the form:

If A then B
A
∴ B

where A and B denote propositions of any degree of complexity. An influential form of proof theory is known as ‘natural deduction’, because it uses formal rules of inference for each sentential connective in a way that renders proofs intuitively easy to understand. Psychologists have similarly proposed a variety of theories based on ‘natural deduction’ in which they postulate formal rules of inference in the mind (see e.g. Osherson, 1976; Rips, 1994; Braine and O’Brien, 1998).

The second method of evaluating arguments is known as ‘model theoretic’ (Jeffrey, 1981). For the sentential calculus, logicians use truth tables and validate arguments by showing that if the premises are true the conclusion must be true. Consider, for example, the following inference:

1. If they are on course then they can see Kalaga.
2. They cannot see Kalaga.
3. ∴ They are not on course.

We can construct a truth table of all possible combinations of truth values of the two *atomic* propositions, i.e. the propositions that do not contain any connective:

| They are on course | They can see Kalaga |
|---|---|
| True | True |
| True | False |
| False | True |
| False | False |

Each row shows a possible combination of truth values, e.g. the third row shows the case where it is false that they are on course and it is true that they can see Kalaga. We can use the premises to eliminate 'models' of possibilities, i.e. rows in the truth table. Premise 1 eliminates the second of the four possibilities, i.e. the second possibility is false given premise 1. Likewise, premise 2 eliminates the first and third possibilities:

| They are on course | They can see Kalaga | |
|---|---|---|
| True | True | Eliminated by premise 2 |
| True | False | Eliminated by premise 1 |
| False | True | Eliminated by premise 2 |
| False | False | |

Hence, only the fourth possibility survives, in which it is false that they are on course. Thus, the conclusion:

> They are not on course.

is *valid* because in any state of affairs in which the premises are true, it is true too. Psychologists have also based theories of reasoning on model-theoretic methods (see e.g. Johnson-Laird and Byrne, 1991; Polk and Newell, 1995). Our aims in what follows are to outline the theory of mental models, to derive a surprising prediction from the theory, and to describe some experiments corroborating it experimentally.

## 2. The mental model theory of reasoning

The model theory of sentential reasoning aims to characterize the deductions of *naive* individuals, that is, those with no training in logic. The theory is similar to the method of truth tables, but diverges from it in several crucial respects (Johnson-Laird and Byrne, 1991). In order to explain the theory, we will use the logical term 'literal' to refer to atomic propositions or their negations. For example, the following exclusive disjunction about a hand of cards:

> Either there wasn′t a king in the hand or else there was an ace in the hand.

contains two literals: 'there wasn't a king in the hand', and 'there was an ace in the hand'. A point to bear in mind, however, is the difference between a sentence or clause and the proposition that it expresses. Most sentences can be used to express many different propositions, e.g. the disjunction above refers to different hands of cards depending on the circumstances of its utterance. It is laborious to keep writing, ‘the proposition expressed by the sentence,’ and so unless the distinction matters we will use ‘assertion’ to refer to sentences or the propositions that they express.

The model theory is based on a fundamental representational principle:

> The principle of truth: individuals tend to minimize the load on working. . . not what is. . . memory by representing explicitly only what is true, and not what is false.

This principle is subtle, and we need to elucidate three points.

The first point is that reasoners are assumed to build models based on all the information in the premises and on their background knowledge, but we focus here on the connectives that can be defined in terms of truth tables, because sentential reasoning hinges on them.

The second point is that the falsity should not be confused with negation. Negation is a grammatical property of a sentence, whereas falsity is a truth value of a proposition. Hence, a negative sentence, such as ‘There wasn’t an ace’, can assert either a true or a false proposition. Of course, the two notions are interrelated because an affirmative assertion is false if and only if its negation is true. The principle of truth postulates that people tend not to represent falsity, but it allows that they will represent negative assertions provided that they are true.

The third point is that the principle of truth applies at two levels for sentential connectives. At the first level, mental models represent only what is true, that is, each model represents a possibility given the premises. Mental models therefore correspond to true rows in a truth table of the premises. At the second level, however, mental models represent only those literals in the premises that are true within each possibility. In sum, mental models represent only true literals within true possibilities. This point can be best explained by way of an example. Consider an *exclusive* disjunction that contains one negative literal and one affirmative literal:

Either there wasn’t a king or else there was an ace.

It has the following truth table, in which we abbreviate the literals:

| Not a king | An ace | Either not a king, or else ace |
|---|---|---|
| True | True | False |
| **True** | False | True |
| False | **True** | True |
| False | False | False |

In contrast, there are just two mental models of the disjunction, and they correspond to the true literals in the true rows of the table (in bold), as shown in the following diagram:

¬King

Ace

Each line denotes a separate model, ‘¬’ denotes negation and so ‘¬King’ denotes a model of the negative literal, ‘there wasn’t a king’, and ‘Ace’ denotes a model of the affirmative literal, ‘there was an ace’. The first model accordingly represents the possibility in which it is true that there wasn't a king, but it does not make explicit that it is false that there was an ace in this possibility. Similarly, the second model represents that it is true that there was an ace, but it does not make explicit that it is false that there wasn't a king in this possibility. Reasoners need to make mental 'footnotes' to capture the information about what is false, e.g. that the first model exhausts the hands in which there wasn't a king and the second model exhausts the hands in which there was an ace. Johnson-Laird and Byrne (1991) used a special notation to represent such footnotes, but we will forego the notation here.

Mental footnotes about falsity can be used to make each model completely explicit if necessary, which can be done by expressing a false affirmative literal as a true negation, and a false negative literal as a true affirmative. The resulting models are *fully explicit*. For example, the fully explicit models of the exclusive disjunction above are:

| | |
|---|---|
| ¬King | ¬Ace |
| King | Ace |

The theory assumes, however, that the mental footnotes about false cases are rapidly forgotten in all but the simplest of cases. There is a premium on truth in human reasoning. What is left out of models is explicit information about possibilities and literals that are *false*.

The principle of truth governs the models of the other sentential connectives. The representation of a conjunction, such as:

There was a king in the hand and there was an ace

is straightforward. It calls for a single completely explicit model of the only true possibility:

| | |
|---|---|
| King | Ace |

An inclusive disjunction, such as:

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

has the following models of the three true possibilities:

| | |
|---|---|
| King | |
| | Ace |
| King | Ace |

Conditionals are notoriously controversial, and there is no consensus about their interpretation. One initially plausible analysis treats them as having a 'defective' truth table in which they have no truth value when their antecedents are false (see e.g. Wason and Johnson-Laird, 1972, p. 90). Thus, a conditional about a specific hand of cards, such as:

If there was a king in the hand then there was an ace

would be true given that there was a king and ace in the hand, false given that there was a king in the hand but not an ace, but it would have no truth value when it is false that there was a king in the hand. This account, however, leads to insoluble problems with biconditionals, such as:

If, and only if, there was a king in the hand, then there was an ace.

This assertion is true given two possibilities: there was a king and an ace in the hand, or neither of them was in the hand; it is false given that there was a king in the hand but not an ace, or there was an ace in the hand but not a king. In other words, the biconditional has a complete truth table, not a defective one. Yet, it can be paraphrased by a conjunction of two conditionals, such as:

If there was a king in the hand then there was an ace;

and if there was an ace in the hand then there was a king.

How can a definite truth table for the biconditional be equivalent to two conditionals with defective truth tables? The answer is: it cannot be equivalent. Suppose that there wasn't a king in the hand and that there wasn't an ace. It follows that neither conditional in the preceding conjunction has a truth value. Yet, the biconditional is true. The notion that a conjunction of two assertions that have no truth value should somehow lead to a true conjunction is a recipe for nonsense.

The model theory is therefore based, not on a defective truth table, but on an idea that captures a similar intuition. A conditional, such as:

If there was a king in the hand then there was an ace

has only one explicit mental model, which represents the possibility in which the antecedent and consequent literals are both true. Individuals do not immediately appreciate the relevance of cases where the antecedent is false. Hence, they defer a detailed representation of these possibilities, which they represent in a wholly implicit model denoted here by an ellipsis:

```
King      Ace
   . . .
```

Again, reasoners need to make a mental footnote about what is false in the wholly implicit model, namely, it is false that there is a king. The representation of a biconditional:

If, and only if, there was a king in the hand then there was an ace

has exactly the same mental models, but now reasoners need to make a footnote that both the king and the ace are false in the implicit model. Our evidence suggests that reasoners soon lose access to mental footnotes, especially in the case of slightly more complex assertions, and so they play no part in the predictions of the present paper.

Over and Evans (1997, p. 268) have pointed out an apparent problem with this account of conditionals. They assert that the model theory takes conditionals of the form, *If A then B,* to be material conditionals and accordingly equivalent to disjunctions of the form, *not-A or B, or both*. There are, however, notorious difficulties for treating conditionals as equivalent to material conditionals. To mention one such problem, which Over and Evans cite, a material conditional supports the following inference ('strengthening the antecedent') as valid:

If A then C
∴ If A and B then C.

In real life, however, there are many inferences of this sort that no-one is likely to accept as valid, e.g.:

If you strike a match then it will light.
∴ If you strike a match and it is wet then it will light.

The solution to these problems is to bear in mind that the antecedent of a conditional sets up a context for the interpretation of the consequent (see Johnson-Laird, 1986, for the following analysis). Indeed, if the context is common knowledge to both speaker and listener, there is no need to assert the antecedent. For example, a mother observing her child about to grab a forbidden cake can assert:

You′ll get into trouble

where the force of the utterance is:

If you take the cake, then you′ll get into trouble.

One constraint on the antecedent of a conditional is therefore that it must have the illocutionary force of an assertion, and so unlike the consequent it cannot ask a question or make a request. And one corollary is that the antecedent often fails to specify the context completely. For example, a conditional such as:

If you put sugar in your tea then it tastes sweet

has an antecedent that only partly describes the context. It has a *ceteris paribus* condition. This condition accommodates the following strengthened antecedent:

If you put sugar and milk in your tea then it tastes sweet,

but it is violated by strengthening the antecedent in the following way:

If you put sugar and diesel oil in your tea then it tastes sweet.

In fact, most sentences give only cues to the situations to which they refer and so they also have *ceteris paribus* conditions. Just as naive individuals baulk at the preceding inference, so too they baulk at certain valid inferences based on disjunctions, such as the following analogue of 'strengthening the antecedent':

Society hostesses put milk in tea or they put sugar in tea.
∴ Society hostesses put milk in tea, or they put diesel oil in tea, or they put sugar in tea.

There is much more that the model theory has to say about conditionals (see e.g. Byrne, 1989), but the studies that follow concern conditionals with antecedents that describe the contexts as completely as necessary. Thus, the antecedent of the following conditional tells the participants all they need to know about the context:

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

Such conditionals *are* akin to material conditionals. Hence, possibilities in which the antecedent is false are consistent with the conditional – they are true possibilities. But, logically-untrained individuals do not normally treat these cases as initially relevant. That is why they are represented by implicit models that have no content (see the ellipsis in the models above).

Table 1 summarizes the mental models for each of the major sentential connectives according to Johnson-Laird et al. (1992). It also shows the fully explicit models in which negations that are true represent the false literals. The relation between the fully explicit models and truth tables is transparent: they map into the true rows in the truth tables for connectives.

Given the principle of truth, how do people envisage the circumstances in which assertions are *false*? The answer is that they must infer these cases from the true cases represented by mental models (and mental footnotes). Thus, given the mental models for the conditional 'If there was a king then there was an ace':

King Ace
. . .

the mental footnote implies that the king cannot occur in the other possibilities represented by the implicit model. It follows that the conditional is false in this case:

King ¬ Ace

Table 1
The sets of mental models for each of the major sentential connectives and their corresponding fully explicit models (as generated by the computer program implementing the model theory at different levels of expertise). Each line represents a separate model: '¬' denotes negation, '...' denotes a model with no explicit content, and 'iff' denotes 'if and only if'

| Connective | Mental models | | 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 |
| Iff A then B | A | B | A | B |
| | . . . | | ¬A | ¬B |

Barres and Johnson-Laird (1997) have shown that individuals do infer false cases from their knowledge of true cases. They and others have also confirmed the preceding account of what falsifies a conditional (see Oaksford and Stenning, 1992; Johnson-Laird et al., 1998a).The mental model theory 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. The process of sentential reasoning consists in constructing models of the premises and ensuring that conclusions are based on them. However, naïve reasoners develop a variety of inferential strategies (see Johnson-Laird et al., 1998b). We cannot illustrate them all, but we will consider how reasoners make a simple deduction based on two premises.

The premises:

| | | |
|---|---|---|
| If A then B: | A | B |
| | . . . | |
| If B then not C: | B | ¬C |
| | . . . | |

yield an integrated set of mental models

| | | |
|---|---|---|
| A | B | ¬C |
| . | . | . |

These models support the conclusion: *If A then not C*. This conclusion is valid, because it holds in the fully explicit models of the premises

| | | |
|---|---|---|
| A | B | ¬C |
| ¬A | B | ¬C |
| ¬A | ¬B | C |
| ¬A | ¬B | ¬C |

Table 2 summarizes the procedures that we have presupposed above for con-structing integrated models for the conjunction of two premises. We make no strong claims that reasoners form separate sets of models and then combine them; they might instead add information from a premise to their existing models. In either case, however, the theory yields the same results. The procedures assume that reasoners consider each pair-wise combination of individual models from the two sets to be combined. Thus, for instance, given the pair of models in the example above

A B and B ¬C

procedure (1) yields the model

A B ¬C

in which there is no need to have duplicate instances of ¬B. The following pair of models contain contradictory elements

A B and ¬B C

and procedure (2) accordingly yields the null model, which is akin to the empty set. When the null model is combined with any other model, procedure (3) yields the null model. Mental footnotes complicate matters. The computer implementation of the theory operates at several levels of expertise depending on how it copes with mental footnotes, but we will explain only the simplest level of performance in which they are forgotten. Thus, when two implicit models are combined, procedure (4) yields an implicit model

. . . and . . . yield . . .

When an implicit model is combined with an explicit model, procedure (5) usually yields the null model

. . . and B C yield null

The exception, which is stated in Table 2, need not detain us because it does not occur in any of the inferences we shall discuss.

Readers who have not encountered the theory before may worry about such concepts as mental footnotes and ellipses. In fact, the theory postulates that individuals normally reason using mental models, but that in simple inferences they can flesh out their models to make them fully explicit. All the predictions in what follows derive from the account summarized in Tables 1 and 2. Theories based on formal rules of inference postulate distinct processes of understanding and reasoning: reasoners first extract the logical form of premises and then reason by manipulating logical forms according to rules of inference. Theories based on mental models make much less of a distinction: the comprehension of the premises yields a set

Table 2
The procedures for forming a conjunction of two sets of models. The procedures apply to each pairwise combination of models from the two sets. Each procedure is presented with an accompanying example. In principle, the procedures take into account mental footnotes, but we have stated their results when the footnotes have been forgotten

| | |
|---|---|
| 1 | For a pair of explicit models, the result conjoins their elements, and drops any duplicates:<br>A B and B C yield A B C |
| 2 | For a pair of models that contain an element and its contradiction, the result is the null model (akin to the empty set):<br>A B and ¬B C yield null |
| 3 | For the null model combined with any sort of model, the result is the null model:<br>null and A B yield null |
| 4 | For a pair of implicit models, the result is an implicit model:<br>. . . and . . . yield . . . |
| 5 | For an implicit model combined with an explicit model, the result is usually the null model:<br>. . . and B C yield null |
| | But if the explicit model contains no item in common with any model in the same set of models in which the implicit model occurs, the result is the explicit model:<br>. . . and B C yield B C |

of mental models, the evaluation of a given conclusion is a process of verifying a description, and the formulation of a conclusion is the description of a set of models. Polk and Newell (1995, p. 563) write of their model-based account, ‘It explains behavior in terms of standard linguistic processes without the need to posit reasoning-specific mechanisms.’ However, we acknowledge that various deductive strategies exist (see Johnson-Laird et al., 1998a), and that they can call for a check that the conclusion holds in the models of the premises.

Evidence for the model theory has been reviewed elsewhere (see e.g. Johnson-Laird and Byrne, 1991). The principle of truth and its account of conditionals has been corroborated in a study in which the participants had to list both the circumstances in which assertions are true and (on separate trials) the circumstances in which they are false (see Johnson-Laird and Barres, 1994). For example, given an inclusive disjunction of the following form:

If A then 2, or if B then 3

the participants tended to list the following possibilities as true cases:

| | | | |
|---|---|---|---|
| A | 2 | | |
| | | B | 3 |
| A | 2 | B | 3 |

They correspond precisely to the mental models of the assertion (ignoring, as the participants do, the implicit model). The participants listed exactly the same possibilities as true for the conjunction:

If A then 2, and if B then 3.

And, again, these possibilities correspond precisely to the mental models of the assertion. In other words, the theory correctly predicts that intelligent adults envisage the same true possibilities for a conjunction and a disjunction of conditionals.

## 3. Illusory inferences: A new prediction of the model theory

The aim of the present section is to explain a surprising prediction of the model theory. In order for readers to have an intuitive grasp of this prediction, we invite them to make the following inference, and to write down their answer for future reference.

Problem 1: Suppose that the following assertions apply to a specific hand of cards:

If there is a king in the hand then there is an ace in the hand,
or else if there is a queen in the hand then there is an ace in the hand.
There is a king in the hand.
What, if anything, follows?

The principle of truth yields an unexpected consequence about such inferences, which the first author discovered by accident in his computer program, written in

LISP, implementing the model theory, including the principles embodied in Tables 1 and 2. It operates at several levels of expertise, depending on whether it builds simple models, or adds mental footnotes, or fleshes out models to make them fully explicit. In debugging the program, the author happened to give it a problem akin to problem 1 above. From its representation of the mental models, it drew a conclusion equivalent to 'There is an ace in the hand'. To his surprise, however, its fully explicit models implied that this conclusion was invalid. This evaluation of invalidity seemed obviously wrong, and so he spent some time searching, in vain, for a bug in the program. Eventually, he checked the problem by hand, and discovered that the program was correct.

If people reason using mental models, then the program predicts that problem 1 should mislead them. Indeed, if you drew the conclusion:

There is an ace in the hand

then you made a fallacious inference. The conclusion seems compelling, but it is invalid. It does not follow from the premises, though the model theory predicts that reasoners will tend to draw it. Our immediate agenda is accordingly to show how the model theory leads to the predicted conclusion, to explain why the conclusion is invalid, and to address the question of what, if anything, does follow from the premises.

Reasoners should envisage the sort of hand of cards shown on the right of Fig. 1 for the conditional, If there is a king in the hand, then there is an ace in the hand. Similarly, they should envisage the sort of hand of cards shown on the left of Fig. 1 for the conditional, If there is a queen in the hand, then there is an ace in the hand. Given an exclusive disjunction of the two conditionals, reasoners should consider that one or other of the two hands must be the case, and so even at this stage they should infer that there is an ace in the hand. The further assertion that there is a king in the hand is categorical, i.e. it picks out the hand on the right in Fig. 1, and so it seems that there must be an ace in the hand.

Of course, reasoners may not form visual images of the two hands. Mental

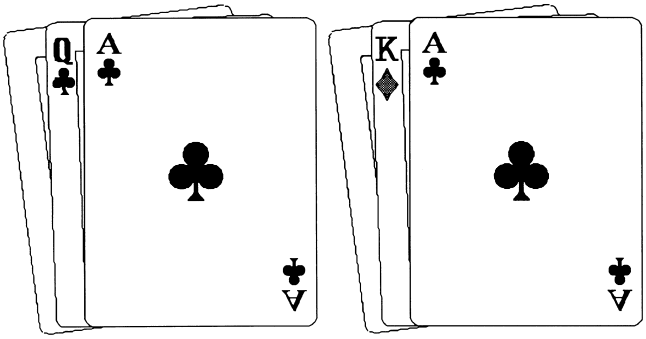

Fig. 1. The two sorts of hands that individuals are likely to envisage in order to represent the assertion: 'If there is a king in the hand then there is an ace in the hand, or else if there is a queen in the hand then there is an ace in the hand'.

models do not necessarily take the form of images. And so we will state the model theory's predictions in general terms. The first conditional in the disjunction elicits the models:

King Ace

. . .

and the second conditional in the disjunction elicits the models:

Queen Ace

. . .

The two conditionals are in an exclusive disjunction. This connective, as Table 1 shows, calls merely for combining the two sets of alternative possibilities. Hence, the mental models of the disjunctive premise as a whole are:

King Ace

Queen Ace

. . .

though individuals are likely to forget the implicit model (denoted by the ellipsis). The categorical premise:

There is a king

describes what is definitely the case, and it eliminates everything except the first model. Hence, the conclusion that seems to follow is:

There is an ace.

An alternative strategy, equally compatible with the model theory, starts with the categorical premise that there is a king, combines it with the first conditional (if there is a king then there is an ace) to infer that there is an ace, and treats the other conditional as irrelevant. Once again, it seems to follow that there is an ace.

This conclusion is compelling, but it is a fallacy granted a disjunction between the two conditional assertions. The proof of the fallacy is simple, and it takes into account that the disjunction could be inclusive or exclusive, and that the conditionals could be one-way conditionals or biconditionals:

1. The disjunction (whether it is inclusive or exclusive) is compatible with the falsity of one of the two conditionals.
2. Therefore, the first conditional could be false.
3. If the first conditional is false, then (whether it is a one-way conditional or a biconditional) one possibility is that there is a king and not an ace. Indeed, naive individuals, as we have seen, generate this possibility when asked to list the false cases of a conditional.
4. Therefore, there is no guarantee that there is an ace even though there is a king.

In other words, the only way in which the conclusion:

There is an ace

could be valid is if there is an interpretation of each conditional that guarantees the presence of an ace even when the conditional is *false*. That is, the following inference would have to be valid:

It is false that if there is a king then there is an ace.
∴ Therefore, there is an ace.

But, that is impossible. Hence, apart from one caveat to which we will return, there is no reasonable interpretation of either the disjunction or the conditionals that yields a valid inference that there is an ace.

Is there any conclusion to problem 1 that *is* valid? If the two conditionals are one-way conditionals about the same hand of cards, then the first conditional is false in case there is a king in the hand and not an ace, and the second conditional is false in case there is a queen in the hand and not an ace. Hence, either way, the following conclusion is valid:

There is *not* an ace in the hand.

In any other case, such as a biconditional interpretation of the conditionals, or an inclusive interpretation of the disjunction, or both, the correct response is that nothing follows from the premises. Any reader who made either of these two responses is to be congratulated – they occur rarely.

In earlier accounts of the model theory, errors were supposed to arise because reasoners overlooked possible models of the premises, and so, for example, Johnson-Laird and Byrne (1991) argued that erroneous conclusions would tend to be consistent with the premises. Illusory inferences such as problem 1 show that this claim was wrong. The mental models of its premises are not a subset of the logically correct models, i.e. the fully explicit models. As we showed earlier, the mental models for the disjunctive premise of problem 1 are as follows (assuming an exclusive disjunction of one-way conditionals):

| | | |
|---|---|---|
| K | | A |
| | Q | A |

. . .

where ‘K’ denotes a model of a king in the hand, and so on. The fully explicit models, however, are as follows:

| | | |
|---|---|---|
| ¬K | Q | ¬A |
| K | ¬Q | ¬A |

where the first model holds if the first conditional is true and the second conditional is false, and the second model holds if the first conditional is false and the second conditional is true. The discovery that mental models could be disjoint with the logically correct models of a premise came as a shock to us.

The possibility of deductive fallacies that are compelling illusions is of intrinsic interest. Yet, if such fallacies occur, they could arise because individuals are reasoning incorrectly, or, as Dan Sperber has noted (personal communication), because individuals misinterpret the premises but otherwise reason correctly. As far as the model theory is concerned, the line between comprehension and reasoning is a fine one, and some theorists have even doubted that it exists (see Polk and Newell, 1995). We make no strong claims either way. Perhaps what is more important is to have a theory of the whole process of reasoning and to determine whether the theory is correct. We therefore needed to test whether the model theory's predictions about illusory inferences were correct. To find more potential illusions, we modified the computer program so that it searched automatically for them amongst a vast set of sentential inferences. By definition, a potential illusion is one in which the mental models of the premises yield a conclusion that differs from the correct conclusion, which is supported by the fully explicit models of the premises. The program revealed that potential illusions should be relatively rare. In most cases, the failure to represent what is false does not lead to a fallacy, i.e. the mental models suffice to yield a valid conclusion. These inferences can serve as control problems for which reasoners should draw the correct conclusions. We designed Experiment 1 to test the difference between illusory inferences and matched controls.

## 4. Experiment 1

### *4.1. Design*

Each participant carried out four problems. The program predicted that two of the problems should yield illusions, and two of them should be control problems. The first potential illusion was:

Illusion 1. Suppose you know the following about a specific hand of cards:

If there is a king in the hand then there is an ace in the hand, or else
if there isn't a king in the hand then there is an ace in the hand.
There is a king in the hand.
What, if anything, follows?

The model theory predicts that reasoners will draw the illusory conclusion: there is an ace in the hand. This conclusion is fallacious, because the first conditional could be false, and so there is no guarantee that there is an ace.

The computer program revealed that a still simpler problem should yield a illusion:

Illusion 2: Suppose you know the following about a specific hand of cards:

If there is a king in the hand then there is an ace, or else there is an ace in the hand.
There is a king in the hand.
What, if anything, follows?

The disjunction yields the following mental models:

Table 3
The two illusions and the two control problems of Experiment 1 (stated in abbreviated form). The table shows for each premise its mental models (on the left) and its fully explicit models (on the right), as generated by the computer program implementing the model theory, which here treats the conditionals as conditionals rather than biconditionals, and ‘or else’ as an exclusive disjunction

| | Illusions | | | | | Control problems | | | |
|---|---|---|---|---|---|---|---|---|---|
| 1 | If king then ace, or else if not king then ace. | | | | 1′ | King and not ace, or else ace and not king. | | | |
| | K | A | K | ¬A | | K | ¬A | K | ¬A |
| | ¬K | A | ¬K | ¬A | | ¬K | A | ¬K | A |
| | . . . | | | | | | | | |
| | There is a king. | | | | | There is a king. | | | |
| | K | A | K | ¬A | | K | ¬A | K | ¬A |
| 2 | If king then ace, or else ace. | | | | 2′ | If king then ace, or else not king. | | | |
| | K | A | ¬K | ¬A | | K | A | K | A |
| | | A | | | | ¬K | | ¬K | A |
| | . . . | | | | | . . . | | K | ¬A |
| | There is a king. | | | | | There is a king. | | | |
| | K | A | | null | | K | A | K | A |

K A
A
. . .

The second premise eliminates all but the first model, from which it follows:

There is an ace.

This conclusion is a fallacy. The conditional in the first premise could be false, and so once again there is no guarantee that there is an ace even though there is a king. In fact, the premises of this problem are self-contradictory, i.e. they yield the null model (see Table 3) from which nothing follows (in contrast to the propositional calculus, which permits the derivation of any conclusion whatsoever from a self-contradiction). When naive individuals do detect a contradiction, they point it out or else say that nothing follows from the premises (Johnson-Laird and Savary, 1996). If the model theory is right, however, naive individuals will fail to notice the contradiction.

The first control problem was:

Control Problem 1′: Suppose you know the following about a specific hand of cards:

> There is a king in the hand and there is not an ace in the hand, or else there is an ace in the hand and there is not a king in the hand.
> There is a king in the hand.
> What, if anything, follows?

The first premise yields the mental models:

| | |
|---|---|
| K | ¬A |
| ¬K | A |

and the categorical premise eliminates the second model. The conclusion:

> There is not an ace

follows validly. The second control problem was:
Control Problem 2′: Suppose you know the following about a specific hand of cards:

> If there is a king in the hand then there is an ace in the hand, or else there is not a king in the hand.
> There is a king in the hand.
> What, if anything, follows?

The mental models of the first premise are:

| | |
|---|---|
| K | A |
| ¬K | |
| . . . | |

and the categorical premise eliminates all but the first model from which it follows validly that there is an ace.

The four problems are summarized in Table 3 together with their mental models and their fully explicit models. The four problems were presented to each participant in a different random order. The participants were asked to state in their own words what, if anything, followed from the premises. They also rated on a five-point scale how confident they were in their conclusions.

### 4.2. *Materials*

Each inference was about a different pair of cards, and no participant encountered the same pair of cards more than once. We made two different allocations of pairs of cards to the four deductions, and half the participants were tested with one allocation and half the participants were tested with the other allocation.

### 4.3. *Procedure*

The participants were tested individually in a quiet room. They were told that they would be asked to make a series of inferences from premises about specific hands of cards, and that each problem concerned a different hand. Their task was to write down what followed from the premises: they should draw a conclusion that must be true about the specific hand of cards given that the premises were true. They were told to respond ‘nothing follows’ if they considered that there was no conclusion that *must* be true given the premises. They had to take into account only the information stated in the premises. They could take as long as they wanted in order to deal with each problem. In addition, they were asked to write down their confidence in their conclusion by assigning it a number on a five-point scale, where 1 signified ‘no confidence' and five signified 'highly confident'. Finally, they were asked if they had any questions about the task. When the experimenter was satisfied that the participants understood their task, he presented the problems one at time. Each problem was typed on a separate sheet of paper, and the participants wrote their response on the sheet.

### 4.4. Participants

Sixteen Princeton students carried out the experiment. None of them had received any training in logic. They were paid $5 for participating in the experiment, which lasted for about 10 min.

### 4.5. Results

Table 4 presents the most frequent conclusion that the participants drew to each pair of premises. As the Table shows, the participants made the fallacious inferences in every case, i.e. 100% of their conclusions to the illusions were the predicted invalid conclusions. In contrast, the participants were nearly always correct in their conclusions for the control problems. Every participant performed more accurately with the control problems than with the illusions ($P = 0.5^{16}$, i.e. less than 1 in 65 000). None of the participants pointed out that illusion 2 had self-contradictory premises. They were highly confident in all of their conclusions, both their invalid conclusions to the illusory problems (a mean confidence rating of 4.4 on a five-point scale) and their valid conclusions to the control problems (a mean confidence rating of 4.3).

### 4.6. Discussion

The illusions have conclusions that seem obvious, that nearly everyone draws, but that are wrong. Unlike traditional logical fallacies, such as *affirmation of the consequent* (If A then B, B, ∴ A) they are compelling both to logically naive individuals and to experts too. We ourselves had at first to keep checking them by referring back

Table 4
The percentages of conclusions that the participants of Experiment 1 drew for themselves, i.e. the percentages of predicted invalid conclusions for the illusions, and the percentages of predicted valid conclusions for the control problems ($n = 16$)

| | | Illusions | | | | Control problems | |
|---|---|---|---|---|---|---|---|
| 1 | | If king then ace, or else | | 1′ | | King and not ace, or else ace | |
| | | if not king then ace. | | | | and not king. | |
| | | There is a king. | | | | There is a king. | |
| | ∴ | There is an ace: | 100 | | ∴ | There is not an ace: | 88 |
| 2 | | If king then ace, | | 2′ | | If king then ace, | |
| | | or else ace. | | | | or else not king. | |
| | | There is a king. | | | | There is a king. | |
| | ∴ | There is an ace: | 100 | | ∴ | There is not an ace: | 100 |

to the output of the computer program. The model theory predicts them on the grounds that reasoners fail to take into account what is false. They are open to some putative alternative explanations, which we will consider in detail later in the paper. However, one obvious possibility is that the participants reasoned carelessly, ignored ‘or else’, and treated the two conditionals as though they were in a conjunction.

In order to test this hypothesis, our colleague Bonnie Meyer carried out a small scale replication of Experiment 1 in which the participants were asked to think aloud as they tackled the problems. She warned one group of participants that the problems could be ‘tricky’ and that they might think that they had reached a correct conclusion when, in fact, they were wrong. Both groups of participants continued to make the fallacies and to get the control problems right. The effect of the warning was certainly to slow the participants down. Here is a typical protocol from a participant tackling a version of illusion 1:

> If there is an ace there is a six or else if there is not an ace there is a six. So there’s always a six. There is an ace.... If there is an ace then there is a six or if there is not an ace then there is a six. So, either way you have a six, and you have an ace. So, you have an ace and a six – I guess. If there is an ace there is a six... or if there is not an ace then there is a six. Yes. So regardless you have a six. I think.... So there has to be something else because that just seems way too easy. And something is wrong somewhere – I don’t know what it is.

Much of the protocol consists in the participant repeating the premises, making the invalid inference, and wondering whether something may have gone wrong. It is not the protocol of someone who is reasoning carelessly, and the use of the phrase ‘either way you have a six’ suggests that the participant had interpreted ‘or else’ as a disjunction, and not as a conjunction. This and other protocols also show that many participants made an immediate inference of the form:

If king then ace, or else if not king then ace.
∴ Ace.

The protocols of the participants who received no warning were typically bare reports of the premises and the conclusion, e.g.:

> If there is a two there is a six or else if there is not two there is a six. There is a two. So there is a six.

In summary, the most likely explanation of our results is that reasoners fail to take into account what is false, and not that the illusions are a result of careless reasoning. Likewise, they do not appear to reflect a misinterpretation of ‘or else’ as ‘and’. Our next experiment was designed to test this conjunctive interpretation more stringently and also to examine another possible explanation.

## 5. Experiment 2

Several expert colleagues, including Lance Rips (personal communication), have suggested that naive reasoners might treat a disjunction of conditionals as though it were an expression in a programming language, such as:

> If there is a king then there is an ace  
> Else if there is not a king then there is an ace.

which validly implies that there is an ace. In other words, they treat the 'or else' as referring only to the antecedents of the two conditionals:

> If there is a king or else there is not a king, then there is an ace.

One might ask why there should be this propensity. If it exists, it could be a further reflection of the model theory. It is not obvious how one can test the hypothesis about the programming-language interpretation, and so we decided to test a more critical prediction for the model theory: the illusions should still occur even when there is an unequivocal exclusive disjunction between the two conditionals *as a whole*. We therefore the expressed the disjunction using the following sort of rubric:

> One of the following assertions is true and one of them is false:  
> If there is a king then there is an ace.  
> If there is not a king then there is an ace.

An exclusive disjunction of the form:

> A or else B

means that one of the two assertions (A,B) is true and one of them is false, and so the *metalinguistic* rubric above is logically equivalent to an exclusive disjunction.

Meta-linguistic reasoning has been investigated using so-called 'knight-and-knave' problems (see Rips, 1989). These problems hinge on the existence of only two sorts of people: knights, who always tell the truth, and knaves, who always lie, e.g.:

> A says, 'I am a knave and B is a knave'.  
> B says, 'A is a knave'.  
> What are A and B?

We wrote a computer program to implement a variety of strategies that naive individuals adopt to solve these problems (see Johnson-Laird and Byrne, 1991). The program contains one component that uses mental models to carry out straightforward sentential deductions, and another component that operates at the metalinguistic level and that calls on the regular deductive machinery as a sub-component. In the same way, a rubric such as:

One of the following assertions is true and one of them is false:

> A  
> B

can be used to call on procedures to construct the models for an exclusive disjunction, A or else B. Likewise, a rubric such as:

If one of the following assertions is true then so is the other:

A
B

can be used to call on procedures to construct the models for a biconditional, if and only if A then B. In the present study, we tested illusion 1 and control problem 1′ from the previous experiment, but we presented only the disjunctive premises, not the categorical premises, i.e. not the assertions about what is definitely the case. We conveyed the disjunction using a metalinguistic rubric, which we can abbreviate as:

One of the following clues is true and one of them is false

where the clues were, in fact, separate conditionals about the same hand of cards. The rubric is an unequivocal exclusive disjunction between the two conditionals as a whole. If readers are skeptical about this claim, then they should ask themselves how else one could convey an exclusive disjunction between two conditionals. The model theory predicts that the participants should still succumb to illusory inferences. If they do, we can be confident that the error is not a result of taking the disjunction as a conjunction.

We lacked the resources for a full-scale experiment, and so we carried out the present study on the World Wide Web. We are probably not the first investigators to carry out an experiment on the Web, but the procedure is unusual, and we should describe the efforts we took to ensure that the results are bona fide. There are three main differences between experiments carried out over the Web and orthodox experiments carried out in the laboratory, and we examine each of them in turn. We also return to the issue of Web experiments in discussing the results of the experiment.

The first difference between orthodox and Web experiments concerns the population of participants. In orthodox experiments, much biographical data is available about the participants, but it is seldom reported or put to use in the present sort of studies. Readers who consult the description of the other experiments in the present paper will discover that the participants were Princeton students who had not previously studied logic. We know almost as much about the participants in the present study. Over a period of many months, the second author had developed a home page consisting of many puzzles and games. This page was visited regularly by people who enjoyed tackling them. We included our experiment on the page with no explicit signs that it was anything but a puzzle for visitors to try to solve. Anyone who looked at the second author’s home page during a period of about two days could participate. The participants are therefore highly likely to come from the population of experienced Web users who enjoy intellectual puzzles. They may have studied logic. If so, as we shall see, it did not help them much.

The second difference concerns the sampling of the participants from their population. In orthodox experiments, investigators control the allocation of participants to conditions and are unlikely to test unwittingly the same individual more than once. In a Web experiment, investigators can exert some control over the allocation of participants to conditions (see Section 5.1), but they cannot prevent an individual from carrying out the same experiment more than once. But, there is no reason why anyone should want to do our experiment more than once. There was no prize for getting the right answer, and the data were anonymous. However, we did raise one obstacle for anyone to participate more than once: they could do so only from different machines (see Section 5.2).

The third difference concerns the control over the participants' performance. In orthodox experiments, the investigator has some control and knowledge of how the participants tackled the inferential problems. In our Web experiment, the participants did not generate their own conclusions but made their response by selecting one of three possible options, a procedure that is convenient for the Web but liable to introduce some noise into the data, because the participants could guess. They could also have used paper and pencil to draw diagrams or to write truth tables. They could have consulted friends or textbooks. A team of participants could have made a single response. There is little motivation in the present case for such excesses, and again if anyone did make such consultations, they do not seem to have helped performance much.

As far as we can determine there are no other significant differences between the two sorts of experiment, the participants are likely to be similarly motivated in both cases. In principle, participants on the Web could be paid for their performance if they were prepared to sacrifice anonymity. Likewise, there is no difference in ethical issues. Our participants signed no consent forms, but there are many other face-to-face empirical studies in which the participants are either deceived or participate without giving their formal consent beforehand. The only deception in our study was that we did not declare that the puzzles were part of an experiment. This degree of deception was considered tolerable by the University's Review Panel, which approved the experiment.

### *5.1. Design and materials*

We tested one group of 20 participants with two problems:
Illusion 1: Suppose you are playing cards with Billy and you get two clues about the cards in his hand. You know that one of the clues is true and that one of them is false, but unfortunately you don't know which one is true and which one is false:

> If there is a king in his hand then there is an ace in his hand.
> If there is not a king in his hand then there is an ace in his hand.
> Please select the correct answer:
> (a) There is an ace in Billy's hand.
> (b) There is not an ace in Billy's hand.
> (c) There may, or may not, be an ace in Billy's hand.

Control Problem 1′: Suppose you are playing cards with Billy and you get two clues about the cards in his hand. You know that one of the clues is true and that one of them is false, but unfortunately you don’t know which one is true and which one is false:

> There is a king in his hand, and there is not an ace in his hand.
> There is not a king in his hand, and there is an ace in his hand.
> Please select the correct answer:
> (a) There is an ace in Billy’s hand.
> (b) There is not an ace in Billy’s hand.
> (c) There may, or may not, be an ace in Billy’s hand.

Table 3 presents the mental models and the fully explicit models for both disjunctions (see the models of the disjunctive premises of illusion 1 and control problem 1′). Half the participants received the illusion and then the control, and half the participants had the two problems in the opposite order.

### 5.2. Procedure and participants

We ran the experiment until 20 Web users had answered the two problems on the second author’s home page. The HTML form automatically forwarded their answers anonymously to the second author’s email address. It also forwarded the address of the machine from which they had responded, and, in order to reduce the probability that one individual carried out the experiment twice, we used the results for the first 20 Web users from different machines.

### 5.3. Results and discussion

The participants succumbed to the illusions: 15 out of the 20 of them inferred that there was an ace, and the remaining five participants chose the indeterminate option. The indeterminate option may reflect one of three possibilities: mere uncertainty, a biconditional interpretation of the conditionals, or an interpretation of the disjunction as somehow inclusive. Fifteen participants also made the correct ‘indeterminate’ response to the control problem. Indeed, the participants performed reliably better with the control problem than with the illusion, albeit the two problems have distinct sorts of correct conclusion (there were 15 participants in accord with the prediction, and five ties, Sign test, $P = 0.5^{15}$, i.e. less than 1 in 30 thousand).

One potential misinterpretation of the premises concerns the phrase in the rubric: ‘You know that one of the clues is true and that one of them is false, but unfortunately you don’t know which one is true and which one is false.’ The participants may have interpreted this phrase to mean that one clue was genuine, i.e. something that is truly a clue, and that the other was either not a genuine clue or one of a doubtful truth value (Dan Sperber, personal communication). They may then have assumed that a genuine clue must have a true antecedent, and so its consequent follows. But, this interpretation of a genuine clue seems implausible. Suppose a detective asserts that he has received a genuine clue to the murderer, namely, ‘if he used gloves, then they will have blood on them.’ We would not ordinarily conclude from such a clue that the murderer used gloves. A conditional clue can be genuine and true without implying that its antecedent is true.

Several individuals have expressed misgivings about carrying out experi ments on the Web (see the Editorial in this issue). We have tried to address the main problems in the introduction to this experiment. We concede that we have little knowledge of the population that we sampled, other than that they are likely to be experienced Web users who enjoy intellectual puzzles. We also concede that we had no control over how the participants tackled the problems. They could have used truth tables to try to work out the correct responses. Yet, they succumbed to the illusory inference and got the control problem right. Bearing in mind the misgivings about the Web, however, we carried out a third experiment, which was designed to outflank any alternative hypothesis based on the interpretation of conditionals.

## 6. Experiment 3

The computer program revealed a set of potential illusions based on disjunctive premises, and the present experiment tested whether the illusions occurred. We examined four problems, and we will explain the predictions for each of them, using the same content for convenience, though in the experiment the contents differed from one problem to another.

The first problem was:

> Only one of the following two assertions is true:
> Albert is here or Betty is here, or both.
> Charlie is here or Betty is here, or both.
>
> This assertion is definitely true:
> Albert isn’t here and Charlie isn’t here.
> What follows?

The model theory predicts that the disjunction of disjunctions yields the following mental models:

| | | |
|---|---|---|
| Albert | | |
| | | Betty |
| Albert | | Betty |
| | Charlie | |
| | Charlie | Betty |

where ‘Albert’ denotes ‘Albert is here’, and so on. The categorical premise eliminates all the models except:

Betty.

Hence, the mental models of the premises predict that participants should conclude:

Betty is here.

This conclusion renders both disjunctive premises true, which contravenes the claim that ‘Only one of the following two assertions is true’. The premises are thus contradictory, but according to the theory, naive individuals will draw the conclusion above without noticing the contradiction. According to the sentential calculus, any proposition whatsoever follows from a contradiction, and so the conclusion above follows from the premises, but so does its negation:

Betty is not here.

According to the mental model theory, however, a contradiction yields the null model, which is akin to the empty set, and nothing follows from the null model (see Table 2). Indeed, as we remarked earlier, naive individuals make a sensible response when they detect a contradiction. They point out that the premises contradict one another or they say that nothing follows from them. We therefore treated either of these responses as correct, and treated reasoners who concluded that Betty is here as having succumbed to an illusion – at the very least, according to the propositional calculus, they have overlooked that they are just as entitled to conclude that Betty is not here.

The second problem was a control:

Only one of the following two assertions is true:

Albert is here or Betty is here, or both.

Charlie is here or Betty is here, or both.

This assertion is definitely true:

Albert isn’t here and Betty isn’t here.

What follows?

The disjunction of disjunctions, as before, yields the mental models:

| | | |
|---|---|---|
| Albert | | |
| | | Betty |
| Albert | | Betty |
| | Charlie | |
| | Charlie | Betty |

The categorical premise eliminates the models containing Albert or Betty, and so all that remains is:

Charlie

which yields the conclusion:

Charlie is here.

This conclusion is valid, and so the failure to take into account what is false should not vitiate the participants' ability to draw this conclusion.

The third problem was an illusory inference based on an exclusive disjunction:

> Only one of the following two assertions is true:
> Albert is here or else Betty is here, but not both.
> Charlie is here or else Betty is here, but not both.

This assertion is definitely true:

> Albert is here and Charlie is here.

Table 5
The four problems in Experiment 3 (stated in abbreviated form). The table shows for each premise the mental models (on the left) and the fully explicit models (on the right)

| | Illusions | | | | | | | Control problem | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Only one true: | | | | | | 2 | Only one true: | | | | | |
| | A or B. | | | | | | | A or B. | | | | | |
| | C or B. | | | | | | | C or B. | | | | | |
| | A | | | A | ¬C | ¬B | | A | | | A | ¬C | ¬B |
| | | | B | ¬A | C | ¬B | | | | B | ¬A | C | ¬B |
| | A | | B | | | | | A | | B | | | |
| | | C | | | | | | | C | | | | |
| | | C | B | | | | | | C | B | | | |
| | Definitely true: | | | | | | | Definitely true: | | | | | |
| | not-A and not-C | | | | | | | not-A and not-B | | | | | |
| | ¬A | ¬C | B | null | | | | ¬A | C | ¬B | ¬A | C | ¬B |
| 3 | Only one true: | | | | | | | | | | | | |
| | A or else B | | | | | | | | | | | | |
| | C or else B | | | | | | | | | | | | |
| | A | C | | A | ¬C | ¬B | | | | | | | |
| | | | B | ¬A | C | B | | | | | | | |
| | | | | A | ¬C | B | | | | | | | |
| | | | | ¬A | C | ¬B | | | | | | | |
| | Definitely true: | | | | | | | | | | | | |
| | A and C | | | | | | | | | | | | |
| | A | C | ¬B | | null | | | | | | | | |
| 4 | Only one true: | | | | | | | | | | | | |
| | A or else B | | | | | | | | | | | | |
| | C or else B | | | | | | | | | | | | |
| | A | C | | A | ¬C | ¬B | | | | | | | |
| | | | B | ¬A | C | B | | | | | | | |
| | | | | A | ¬C | B | | | | | | | |
| | | | | ¬A | C | ¬B | | | | | | | |
| | Definitely true: | | | | | | | | | | | | |
| | not-A and not-C | | | | | | | | | | | | |
| | ¬A | ¬C | B | | null | | | | | | | | |

The disjunction of disjunctions yields the following mental models:

```
Albert     Charlie
                      Betty
```

The categorical premise asserts the first of these models, and so the participants should tend to draw the conclusion:

Betty is not here.

In fact, the categorical premise itself renders both disjunctive premises true and contravenes the claim that only of them is true.

The fourth problem was also based on an exclusive disjunction:

Only one of the following two assertions is true:
  Albert is here or else Betty is here, but not both.
  Charlie is here or else Betty is here, but not both.
This assertion is definitely true:
  Albert isn't here and Charlie isn't here.

The categorical premise eliminates the model:

```
Albert          Charlie
```

Table 6
The percentages of illusory and correct conclusions that the participants drew for themselves in Experiment 3 ($n = 40$)

| | Illusions | | | Control problem | |
|---|---|---|---|---|---|
| 1 | Only one true: | | 2 | Only one true: | |
| | A or B. | | | A or B. | |
| | C or B. | | | C or B. | |
| | Definitely true: | | | Definitely true: | |
| | not-A and not-C | | | not-A and not-B | |
| | Fallacious conclusion: ∴B: | 85 | | | |
| | Correct: Nothing follows: | 15 | | Correct conclusion: ∴C: | 98 |
| 3 | Only one true: | | | | |
| | A or else B. | | | | |
| | C or else B. | | | | |
| | Definitely true: | | | | |
| | A and C | | | | |
| | Fallacious conclusion: ∴B: | 63 | | | |
| | Correct: Nothing follows: | 28 | | | |
| 4 | Only one true: | | | | |
| | A or else B. | | | | |
| | C or else B. | | | | |
| | Definitely true: | | | | |
| | not-A and not-C | | | | |
| | Fallacious conclusion: ∴B: | 68 | | | |
| | Correct: Nothing follows: | 28 | | | |

and so participants should draw the conclusion corresponding to the remaining model:

Betty is here.

Once again, however, its truth renders both disjunctions true.

### 6.1. Design and procedure

Each participant carried out the same four inferences, which are presented in Table 5 with their mental models and their fully explicit models. The problems were presented in the same order as in the table on a single sheet of paper. The participants, who were tested individually, were asked to write in their own words, what, if anything, followed from each set of premises. They were told to write, ‘nothing follows,’ if they considered that there was no conclusion that must be true given the premises. They were not allowed to go back to change their answers to earlier problems.

The experiment was carried out by a set of 22 naive experimenters who did not know anything about the predictions of the model theory. They were participants in a course on Cognitive Science at Princeton University, and they were asked to give the problems to one or two fellow students. We used a fixed order of presentation to keep the instructions to the experimenters as simple as possible.

### 6.2. Participants

Forty students at Princeton University took part voluntarily in the experiment. None of them had received any training in logic.

### 6.3. Results and discussion

Table 6 presents the percentages of the conclusions that the participants drew for themselves. We explained earlier that the illusions are self-contradictory, and that we accordingly treated the response ‘nothing follows’ as correct. As the table shows, the participants’ responses corroborated the model theory. Of the 40 participants, 33 drew one or more illusory conclusions but were correct on the control problem, and the remaining seven participants were ties (Sign test, $P = 0.5^{33}$, i.e. less than 1 in 8 billion).

In general, the participants succumbed to the illusions but drew the correct conclusion to the control problem. They may have been less susceptible to the illusions based on exclusive disjunctions than to the illusion based on the inclusive disjunction, but since they always tackled the exclusive disjunctions last, they may have acquired some expertise with the task during the course of the experiment. From their introspective reports, only a few participants (among those who responded ‘nothing follows’) realized that the premises of the illusions were self-contradictory.

No account of the interpretation of conditionals can illuminate these results, because the problems do not contain any conditionals. Similarly, the hypothesis that reasoners treat disjunctions of conditionals as expressions in a programming language does not apply to these results, either. For example, problem 1 is akin to the following expression in a Lisp-like language:

```
(and (or-exclusive
(or-inclusive (A is true)(B is true))
(or-inclusive (C is true)(B is true)))
(and (A is false)(C is false)))
```

The evaluation of this expression, however, yields the correct response that the expression as a whole is inconsistent, i.e. it is false. The analogy to a programming language accordingly fails to account for the illusion.

The disjunction of the two premises in the present experiment was conveyed by the assertion: ‘Only one of the following assertions is true’. It is difficult to believe that this form of words is regularly taken to mean that both of the assertions are true, the interpretation required for a conjunction of the two premises. Hence, the experiment also casts doubt on the hypothesis that individuals treat exclusive disjunction as though it were a conjunction.

## 7. General discussion

A surprising, and only recently discovered, consequence of the theory of mental models is its prediction of illusory inferences, that is, inferences that lead to compelling but fallacious conclusions. For example:

> Illusion 1. Suppose you know the following about a specific hand of cards:
> If there is a king in the hand then there is an ace in the hand, or else
> if there isn’t a king in the hand then there is an ace in the hand.
> There is a king in the hand.

Given these premises, all the participants in Experiment 1 drew the conclusion:

> Therefore, there is an ace in the hand.

They were highly confident that it was correct, as were we when we first tackled the inference. We have also observed the same response informally, only one person among the many distinguished cognitive scientists to whom we have given the problem made a correct response. Several hundred individuals at public lectures from Stockholm to Seattle have made the same error, just one person asserted that nothing followed from the premises. And less than 2% of nearly a thousand prospective students at the University of Padua avoided the error (Vittorio Girotto, personal communication). Yet, the conclusion that there is an ace is invalid.

Experiment 2 showed that a rubric that made explicit an exclusive disjunction of conditionals also led to an illusion:

> Illusion 2. Suppose you are playing cards with Billy and you get two clues about the cards in his hand. You know that one of the clues is true and that one of them is false, but unfortunately you don't know which one is true and which one is false:
>
> If there is a king in his hand then there is an ace in his hand.
> If there is not a king in his hand then there is an ace in his hand.

The majority of participants inferred invalidly that there was an ace in the hand.

The model theory also predicts that illusions should occur with some premises based on disjunctions, and Experiment 3 corroborated their existence with, for example, the following problem:

> Illusion 3. Only one of the following two assertions is true:
> Albert is here or Betty is here, or both.
> Charlie is here or Betty is here, or both.
> This assertion is definitely true:
> Albert isn't here and Charlie isn't here.
> What follows?

Most of the participants drew the illusory conclusion:

> Betty is here.

Our results raise several issues that we need to examine, though their answers tend to overlap. The first issue is whether there is any way in which the conclusions to the illusions could be valid. For instance, could the conclusion to illusion 1 above:

> There is an ace in the hand

be valid? In fact, the conclusion is invalid given that the two conditional premises occur in a disjunction, whether it is inclusive or exclusive, and whether they are one-way conditionals or biconditionals. A disjunction implies that one of the conditionals could be false, and so even though there is definitely a king, there is no guarantee that there is an ace, because the first conditional could be false. Similarly, the same conclusion to illusion 2 is also invalid. The exclusive disjunction in this case is unequivocal, and so one of the conditionals is definitely false. The falsity of either one of them guarantees that there is not an ace in the hand. Only the conclusion to illusion 3 is arguably valid, because the premises are self-contradictory, and, according to logic, self-contradictions validly yield any conclusion whatsoever. The participants, however, tended not to notice the self-contradiction, but drew the conclusion unaware that its negation also followed logically from the premises. In summary, the conclusions *are* illusions.

The second issue is what conclusions, if any, are valid for the illusory problems. The correct response to illusions 1 and 2 depends on the interpretation of the premises. The most likely interpretation of the conditionals is that they are 'one way', not biconditionals, and that the disjunction is exclusive. In this case, the valid conclusion is:

There is *not* an ace in the hand.

In any other case, such as a biconditional interpretation of the conditionals, an inclusive interpretation of the disjunction, or an interpretation in which the conditionals range over a population of hands, the only correct response is that nothing follows from the premises.

The third, and most important, issue is whether there is any alternative explanation for the illusions apart from the mental model theory. Because so many experts have succumbed to illusions, we have accumulated a list of potential alternative explanations, which we will review. There are five principal hypotheses.

(1) *The task, instructions, or premises of the illusions are so complex, ambiguous, or pragmatically odd, that they confuse the experimental participants, who therefore succumb to the illusions*. This vein of criticism is encapsulated in the following remarks made by an expert: ‘The premises are multiply ambiguous, and the inferential task lacks a clear pragmatic context. The inferences don’t make sense.’ Experiment 1 raised several problems for such objections. If the participants had been confused about the illusions, then they would have lacked confidence in their conclusions. In fact, they were highly confident in both their illusory conclusions and their valid control conclusions, with no reliable difference between these confidence ratings. They performed very well with the control problems, which are of a similar degree of surface complexity. For instance, the only difference between illusion 2 and control problem 2′ is that the premises of the illusion contain the literal, ‘there is an ace’ in place of the control problem’s literal, ‘there is not a king’ (see Table 3). Yet the participants were 100% wrong with the illusion and 100% right with the control problem. It is hard to see how an affirmative literal in place of a negative one could transform a problem into one that is so complex, ambiguous, or pragmatically odd, that it confused the participants. Moreover, confusion in reasoning normally leads to a variety of different conclusions – as in the case, say, of difficult syllogisms, but in our experiment everyone drew one and the same erroneous conclusion. Precise pragmatic theories may have much to contribute to our understanding to illusory inferences (cf. Sperber et al., 1995), but vague appeals to a lack of clear ‘pragmatics’ or clear premises do not illuminate our results. They fail to explain the excellent performance with the control problems and the high level of confidence in both the fallacies and the control conclusions. The illusions are deceptively simple, not confusing. The ultimate refutation of any alternative hypotheses based on some feature of the premises, task, or instructions, is a recent experiment in which the illusions and control problems were based on the *same* premises (Johnson-Laird and Goldvarg, 1997). The problems differed only in the questions that the participants had to answer, which were all conjunctions, e.g.:

One of the following premises is true and one is false:
There is a king in the hand, or an ace, or both.
There is a queen in the hand and there is an ace.
Is it possible that there is a queen and an ace in the hand?

Given this problem, nearly everyone responded invalidly, ‘yes’. Yet, the presence of a queen and an ace renders both of the premises true, contrary to the rubric that only one of them is true. Given the same premises but the different control question (on a separate trial with different content):

> Is it possible that there is a king and an ace in the hand?

nearly everyone responded correctly, ‘yes’. It is unlikely that the difference between the two questions could have yielded large pragmatic effects, and in any case the question eliciting the fallacy occurred as a control question in other problems.

2. *The participants failed to notice the disjunctions of premises in the illusions, and treated them as conjunctions*. This hypothesis seems unlikely, because it would convert some of the control problems into self-contradictions (e.g. problem 1′ in Experiment 1). Skeptics, however, have continued to defend this alternative explanation on the grounds that ‘or else’ could be interpreted as ‘and’ unless the interpretation yields a contradiction. In our view, the instructions to the participants made the disjunctions completely clear, as did the metalinguistic rubrics in Experiments 2 and 3. But, the hypothesis is decisively eliminated by the ‘think aloud’ protocols in Bonnie Meyer’s study in which the participants treated the disjunctions as disjunctions, and yet still succumbed to the illusions.

3. *The illusions arise because conditionals have an interpretation distinct from the one postulated by the model theory*. There are several versions of this hypothesis: conditionals are interpreted as generalizations about open-ended populations of hands, as having a ‘defective’ truth table, as akin to expressions in a programming language, or as having some other, as yet unknown, sophisticated meaning.

> The interpretation of the conditional:
> If there is a king in the hand then there is an ace in the hand

as an assertion about a set of hands of cards as opposed to a single hand is implausible for Experiment 2, and in any case fails to explain the illusions. The falsity of such a conditional implies that there may be hands in which there are kings without aces, and therefore fails to explain the illusory conclusion.

The hypothesis that conditionals are treated as having a ‘defective’ truth table leads, as we showed earlier, to insoluble problems with biconditionals. But, it can provide an account of illusion 1 if it is combined with some further assumptions. Given illusion 1, which we here abbreviate:

> If king then ace, or else if not king then ace.
> King

reasoners suppose that the presence of the king (as asserted in the categorical premise) implies that the second conditional has no truth value because its antecedent is false. They then assume that the first conditional must be true, and so it follows that there is an ace. However, they have no justification for assuming that the presence of the king guarantees that the first conditional is true. An inference of the form:

There is a king.
∴ If there is a king then there is an ace.

is a license to deduce any consequence whatsoever, because the premise and conclusion in turn imply that the consequent is true, i.e. that there is an ace. We doubt whether anyone would infer that the truth of a conditional's antecedent implies that the conditional as a whole is true. But, the notion of a defective truth table is not so too distant from the model theory itself. It differs only in its unequivocal attachment to the idea that conditionals are discounted when their antecedents are false, whereas the model theory postulates merely that individuals are not initially aware that such cases are consistent with the truth of the conditional.

The disjunction of conditionals in illusion 1 might have a programming language interpretation (L. Rips, personal communication):

If there is king then there is an ace
Else if there is not a king then there is an ace.

Equivalently, reasoners may have treated 'or else' as applying only to the antecedents of the conditionals:

If there is a king or else if there isn't a king, then there is an ace.

The hypothesis would explain the illusory conclusion in Experiment 1, but it is hard to maintain when the disjunction is expressed metalinguistically using the rubric of Experiment 2:

One of the clues is true and that one of them is false.

This assertion clearly states that one conditional as a whole is true and one conditional as a whole is false. If not, then how else could one convey this information?
Some critics have argued that conditionals must have an interpretation that somehow justifies the illusory conclusions, though they have stopped short of providing it. Any such account were it to be forthcoming still fails to offer an explanation of the illusions, such as illusion 3, that depend solely on disjunctions.

4. *Reasoners fail to represent explicitly those propositions that are not stated explicitly in the premises*. This hypothesis, which is a variant of the model theory, seems plausible at first sight. An exclusive disjunction, A or else B, contains two explicit assertions, A and B, and so reasoners represent only these two assertions. This account, however, does not explain the effects of different sentential connectives. Thus, a conjunction, A and B, and inclusive disjunction, A or B or both, and a conditional, if A then B, each contain only two explicit assertions, but their mental models are quite different (see Table 1), and it is these models that predict the illusions.

5. *If reasoners think about one disjunct then they forget about the other*. This hypothesis, which was drawn to our attention by Robert Mackiewicz and Walter Schaeken (personal communication), is yet another variant on the model theory. What the model theory postulates is that people represent only what is true. This principle has the emergent property for exclusive disjunctions that when people think about the truth of one disjunct they forget about the falsity of the other disjunct. But, it is important to realize that the principle of truth underlies what is going on. If reasoners merely forgot this or that clause, then their performance would not be affected by the sentential connective, i.e. they would be liable to forget A or to forget B in dealing with each of the following assertions:

A and B
If A then B
A if and only if B

However, consider the following potent illusion based on a biconditional relation between two premises (Johnson-Laird and Savary, 1996):

If one of the following assertions is true then so is the other:
There is a king if and only if there is an ace.
There is a king.
Which is more probable the king or the ace?

Most of the participants (90%) responded that the two cards are equally probable, which is a fallacy. Both assertions could be false, in which case there is not a king, but there is an ace. The source of the illusion is not that when reasoners think about one assertion, A, they forget about the other, B, but rather that they think about the joint truth of A and B, but not their joint falsity.

Overall, the model theory gives a better account of the phenomena than the rival hypotheses. It predicts all the phenomena, whereas they are post-hoc and account, at best, for only some of the illusions. Unlike the rival hypotheses, the model theory also predicts that any manipulation that emphasizes falsity should reduce the illusions. Recent studies have corroborated this prediction. For example, the rubric, ‘Only one of the following two premises is false,’ reliably reduced the illusions (Tabossi et al., 1998), as did the prior production of false instances of the premises (see Newsome and Johnson-Laird, 1996; and an unpublished study by Vittorio Girotto). Yang and Johnson-Laird (1998a) have shown that illusory inferences occur with quantified assertions. Given, say, the problem:

Only one of the following statements is true:
Not all the plastic beads are red, or
None of the plastic beads are red.
Is it possible that none of the red beads are plastic?

most people answer incorrectly ‘yes’. Two experiments progressively eliminated the fallacies by using instructions designed to overcome the bias towards truth (Yang and Johnson-Laird, 1998b). Performance with the fallacies and the control problems did not differ when the participants were taught to check whether each conclusion was consistent with (1) the truth of the first premise and the falsity of the second premise, and (2) the truth of the second premise and the falsity of the first premise. A final advantage of the model theory is that it bases its predictions on a single principle: reasoners take into account truth, not falsity. In contrast, each of the alternatives has a very limited purview. Occam's razor cuts them off, leaving in place the model theory.

The illusions are an important shortcoming in human reasoning, and they are worth investigating further for their own intrinsic interest. They also allow us to draw some general morals.

A number of other well-established inferential phenomena appear to be special cases of the same mechanism underlying the illusions. For example, inferences of the form known as *modus tollens* are often difficult for naive individuals (see Evans et al., 1993), who fail to draw the valid conclusion, e.g.:

If they are on course then they can see Kalaga.
They cannot see Kalaga.
∴ They are not on course.

They may be difficult because models of conditionals do not normally represent possibilities in which the antecedent of a conditional is false. Likewise, Wason's selection task may be just another way of demonstrating the failure to cope with false possibilities. In this task, participants are asked which of four cards:

A B 2 3

they need to turn over in order to find out whether a conditional assertion is true or false:

If there is an 'A' on one side of a card then there is a '2' on the other side of the card

where each card has a letter on one side and a number on the other side. Their characteristic error is one of omission: reasoners fail to select the '3' card (see e.g. Wason and Johnson-Laird, 1972). The explanation may be that people construct models of what is true, but not of what is false. Hence, any manipulation that helps them to consider false instances of the rule, should improve performance. Several recent studies bear out this explanation (see e.g. Green and Larking, 1995; Liberman and Klar, 1996; Love and Kessler, 1995; and Sperber et al., 1995).

The illusions contravene all current formal rule theories (e.g. Rips, 1994; Braine and O'Brien, 1998). These theories rely solely on valid rules of inference, and so the only systematic conclusions that they can account for are valid ones. The theories therefore need to be amended, either in their implementation or in a more radical way in order to account for the illusions. One idea to save formal rules is that reasoners misapply a suppositional strategy (L. Bonatti and D. O'Brien, personal communications). Unfortunately, this hypothesis makes the wrong predictions in certain cases. Given premises of the following form from the studies carried out by

Yang and Johnson-Laird (1998a,b):

> Only one of the following statements is true:
> Some of the A are not B, or
> None of the A are B.
> Is it possible that none of the B are A?

the majority of participants responded ‘yes’, though the response is a fallacy. A misapplication of a supposition could explain the phenomenon. A supposition of the second premise, None of the A are B, yields the conclusion, None of the B are A, and hence a ‘yes’ response. But, now consider an inference that has the same premises as above, but the question:

> Is it possible that some B are A?

This putative conclusion cannot be validly inferred from a supposition of either premise, and so reasoners should respond ‘no’ to the question. In fact, they correctly respond, ‘yes’. The problem is a control problem, but the misapplication of the suppositional strategy should turn it into an illusion. In other words, a misapplication of rules for suppositions obliterates the distinction between illusory and control problems. There may not be any simple modification to current formal rule theories that can save the phenomena.

Finally, the illusions are relevant to the general debate about human rationality. Some authors argue that errors in reasoning do not occur (see Henle, 1962). Others argue that humans have adapted to be rational, and that apparent errors can often be explained as a rational performance of a different task from the one the experimenter set the participants (see e.g. Anderson, 1990; Oaksford and Chater, 1994; Nickerson, 1996). Still others, as our epigraph from Oscar Wilde illustrates, take for granted human irrationality, at least in the laboratory (Evans and Over, 1996; Over and Evans, 1997). Our view is that human beings are rational in principle, because they can see the force of counterexamples – models that refute conclusions, and because some of them can construct logic and other systems that provide the standards of rationality. Human reasoners can make valid deductions, and they can realize that they have made a valid deduction. But, they can also make errors, and they can realize that they have erred. The illusions can be described as a result of a systematic misunderstanding of premises, but understanding itself has been described as the heart of reasoning (Polk and Newell, 1995). All of the individuals who committed a fallacy when we gave them an illusory inference informally have eventually understood their error.

Are the illusions merely a laboratory phenomenon that depends on weird materials? It seems not, because they also occur in daily life. So persuasive are they, however, that their occurrence probably goes unnoticed by both their perpetrators and their interlocutors. Once again, we turned to the World Wide Web, and searched for the sequence of words, ‘or else if’; the search turned up several illusory inferences. A professor cautioned students, for example:

> ...either a grade of zero will be recorded if your absence [from class] is not

> excused, or else if your absence is excused other work you do in the course will count...

The mental models of this assertion yield the two possibilities that presumably he and his students had in mind:

| | |
|---|---|
| ¬excused | zero-grade |
| excused | other-work-counts |

But, an assertion of the form:

> B if not A, or else if A then C

has very different fully explicit models. Indeed, what the professor should have asserted is, not a disjunction, but a conjunction of the two conditionals:

> A grade of zero will be recorded if, and only if, your absence is not excused, *but* if and only if your absence is excused other work you do in the course will count.

Given the limitations of human working memory, reasoners cannot cope with fully explicit models, still less with complete truth tables. The principle of taking into account what is true and foregoing what is false is a sensible compromise. Truth is more useful than falsity. Just occasionally, however, truth alone leads human reasoners into the illusion that they grasp a set of possibilities that is in fact beyond them.

## Acknowledgements

We thank the members of our laboratory, who have provided much helpful advice: Patricia Barres, Victoria Bell, Zachary Estes, Yevgeniya Goldvarg, Mary Newsome, Kyung-Soo Do, Lisa Torreano, Isabelle Vadeboncoeur, and Yingrui Yang. We thank Bonnie Meyer for running the ‘think aloud’ study. We are also grateful to several Italian colleagues: Patrizia Tabossi collaborated in a study of the illusions, Vittorio Girotto carried out two independent studies of them, Monica Bucciarelli investigated the construction of counterexamples, and Paolo and Maria Sonino Legrenzi collaborated in a study of illusory inferences about probabilities. We thank Kent Bach, Jerry Bruner, Ruth Byrne, Jean-Paul Caverni, Jack Gelfand, Sam Glucksberg, Tony Greenwald, Geoff Hinton, Danny Kahneman, Joel Lachter, Alan Leslie, Rick Lewis, Robert Mackiewicz, Mike McCloskey, George A. Miller, Zenon Pylyshyn, Lance Rips, Walter Schaeken, Eldar Shafir, and the late Amos Tversky, for their helpful suggestions and alternative hypotheses. We also thank Keith Holyoak for drawing our attention to the ‘defective’ truth table and for many cogent comments on earlier versions of this paper. And we thank Jonathan Evans, Dan Sperber, and an anonymous referee, for their useful criticisms of previous drafts of the paper. The research was supported in part by Fonds pour la Formation de Chercheurs et l’aide a la Recherche (Quebec). We presented some of the research to a meeting of SPUDM-15 (Subjective Probability, Utility, and Decision

Making) in Jerusalem, August 1995, and we are grateful to Maya Bar-Hillel and many other participants for their comments.

## References

Anderson, J.R., 1990. The Adaptive Character of Thought. Lawrence Erlbaum Associates, Hillsdale, NJ.

Barres, P.E., Johnson-Laird, P.N., 1997. Why is hard to imagine what is false? Proceedings of the Nineteenth Annual Conference of the Cognitive Science Society, 859.

Braine, M.D.S., O'Brien, D.P. (Eds.), 1998. Mental Logic. Lawrence Erlbaum Associates, Mahwah, NJ.

Byrne, R.M.J., 1989. Suppressing valid inferences with conditionals. Cognition 31, 61–83.

Evans, J.St.B.T., Newstead, S.E., Byrne, R.M.J., 1993. Human Reasoning: The Psychology of Deduction. Lawrence Erlbaum Associates, Hillsdale, NJ.

Evans, J.St.B.T., Over, D.E., 1996. Rationality and Reasoning. Psychology Press, Hove.

Green, D.W., Larking, R., 1995. The locus of facilitation in the abstract selection task. Thinking and Reasoning 1, 121–200.

Henle, M., 1962. The relation between logic and thinking. Psychological Review 69, 366–378.

Jeffrey, R., 1981. Formal Logic: Its Scope and Limits, 2nd edn. McGraw-Hill, New York.

Johnson-Laird, P.N., 1986. Conditionals and mental models. In: Traugott, E.C., ter Meulen, A.,Reilly, J.S., Ferguson, C.A. (Eds.), On Conditionals. Cambridge University Press, Cambridge, pp. 55–75.

Johnson-Laird, P.N., Barres, P.E., 1994. When 'or' means 'and': a study in mental models. Proceedings of the Sixteenth Annual Conference of the Cognitive Science Society. Lawrence Erlbaum Associates, Hillsdale, NJ, pp. 475–478.

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

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

Johnson-Laird, P.N., Goldvarg, Y., 1997. How to make the impossible seem possible. Proceedings of the Nineteenth Annual Conference of the Cognitive Science Society. Lawrence Erlbaum Associates, Mahwah, NJ.

Johnson-Laird, P.N., Legrenzi, P., Girotto, V., Legrenzi, M. S., Caverni, J.-P., 1998. Naive probability: a model theory of extensional reasoning. Psychological Review 106, 62–88.

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

Johnson-Laird, P.N., Savary, F., Bucciarelli, M., 1998. Strategies and tactics in reasoning. In: Schaeken, W. (Ed.), Strategies in Reasoning. Lawrence Erlbaum Associates, Hillsdale, NJ.

Liberman, N., Klar, Y., 1996. Hypothesis testing in Wason's selection task: social exchange cheating detection or task understanding. Cognition 58, 127–156.

Love, R.E., Kessler, C.M., 1995. Focusing in Wason's selection task: content and instruction effects. Thinking and Reasoning 1, 153–182.

Newsome, M.R., Johnson-Laird, P.N., 1996. An antidote to illusory inferences? In: Cottrell, G.W. (Ed.), Proceedings of the Eighteenth Annual Conference of the Cognitive Science Society. Lawrence Erlbaum Associates, Mahwah, NJ, p. 820.

Nickerson, R.S., 1996. Hempel's paradox and Wason's selection task: logical and psychological puzzles of confirmation. Thinking and Reasoning 2, 1–31.

Oaksford, M., Chater, N., 1994. A rational analysis of the selection task as optimal data selection. Psychological Review 101, 608–631.

Oaksford, M., Stenning, K., 1992. Reasoning with conditionals containing negated constituents. Journal of Experimental Psychology: Learning, Memory, and Cognition 18, 835–854.

Osherson, D., 1976. Logical Abilities in Children, Vols. 1–4. Lawrence Erlbaum Associates, Hillsdale, NJ.

Over, D.E., Evans, J.St.B.T., 1997. Two cheers for deductive competence. Current Psychology of Cognition 16, 255–278.

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

Rips, L.J., 1989. The psychology of knights and knaves. Cognition 31, 85–116.

Rips, L.J., 1994. The Psychology of Proof. MIT Press, Cambridge, MA.

Sperber, D., Cara, F., Girotto, V., 1995. Relevance theory explains the selection task. Cognition 52, 3–39.

Tabossi, P., Bell, V.A., Johnson-Laird, P.N., 1998. Mental models in deductive, modal, and probabilistic reasoning. In: Habel, C., Rickheit, G. (Eds.), Mental Models in Discourse Processing and Reasoning. North-Holland, Amsterdam.

Wason, P.C., Johnson-Laird, P.N., 1972. The Psychology of Deduction: Structure and Content. Harvard University Press, Cambridge, MA.

Yang, Y., Johnson-Laird, P.N., 1998a. Illusions in quantified reasoning: how to make the impossible seem possible, and vice versa. Memory and Cognition, (in press).

Yang, Y., Johnson-Laird, P.N., 1998b. Systematic fallacies in quantified reasoning and how to eliminate them (submitted).