THINKING AND REASONING, 2000, *6* (4), 313–333

An antidote to illusory inferences

Carlos Santamaría
Universidad de La Laguna, Tenerife, Spain

P.N. Johnson-Laird
Princeton University, USA

The mental model theory predicts that reasoners normally represent what is true, but not what is false. One consequence is that reasoners should make "illusory" inferences, which are compelling but invalid. Three experiments confirmed the existence of such illusions based on disjunctions of disjunctions. They also established a successful antidote to them: Reasoners are much less likely to succumb to illusions if the inferences concern disjunctions of physical objects (alternative newspaper advertisements) rather disjunctions of the truth values of assertions. The results shed light both on the cause of the illusions and on the current controversy among different theories of reasoning.

INTRODUCTION

Mental models are psychological representations of real, hypothetical, or imaginary situations. They can be used to reason deductively by ensuring that a conclusion holds in all the models of the premises (Johnson-Laird & Byrne, 1991). Each mental model represents a single possibility, capturing what is common to all the different ways in which it might occur. Hence, the theory provides a unified account of reasoning about what is necessary, probable, and possible. A conclusion is necessary—it *must* be true—if it holds in all of the models of the premises, it is probable—it is *likely* to be true—if it holds in most of the models of the premises, and it is possible—it *may* be true—if it holds in at least one of the models of the premises. The theory is accordingly an alternative to the view that deductive reasoning depends on formal rules of inference akin to

Correspondence should be sent to P.N. Johnson-Laird, Department of Psychology, Princeton University, Princeton, NJ 08544, USA. Email: phil@princeton.edu

We thank Bonnie Meyer for running the "think aloud" study. We are grateful for help from several other colleagues: Victoria Bell, Monica Bucciarelli, Ruth Byrne, Yevgeniya Goldvarg, Vittorio Girotto, Paolo Legrenzi, Maria Sonino Legrenzi, Mary Newsome, Fabien Savary, Patrizia Tabossi, and Yingrui Yang. We thank Ruth Byrne, Jonathan Evans, David Green, and David Over for their helpful criticisms of an earlier draft.

http://www.tandf.co.uk/journals/pp/13546783.html

those of a logical calculus (e.g., Braine & O'Brien, 1998; Rips, 1994). The distinction between the two sorts of theories parallels the one in logic between proof-theoretic methods based on formal rules and model-theoretic methods based, say, on truth tables.

The controversy between models and rules has been long but fruitful: It has led to better experiments and to explicit theories implemented in computer programs. Which psychological theory provides a better account of naive human reasoning is controversial, but the experiments have corroborated several tell-tale signs of the use of mental models (see Evans, Newstead, & Byrne, 1993).

First, the greater the number of models that an inference elicits, the harder reasoning is (e.g., Johnson-Laird & Byrne, 1991). Second, the greater the complexity of individual models, the poorer performance is (e.g., Schaeken & Johnson-Laird, 2000). Third, reasoners tend to focus on a subset of the models of the premises, and so erroneous conclusions tend to correspond to such subsets, often just a single model of the premises (e,g., Bauer & Johnson-Laird, 1993). Fourth, procedures for reasoning with mental models rely on counterexamples to refute invalid inferences; they establish validity by ensuring that a conclusion holds over all or the models of the premises (e.g., Bucciarelli & Johnson-Laird, 1999). These procedures can be implemented in a formal system, but current psychological theories (and most AI programs) based on formal rules do not use them. Fifth, it is easier to establish that a situation is possible (only one model needs to be considered) than that it is impossible (all models need to be considered), whereas it is easier to establish that it is not necessary (only one model needs to be considered) than that it is necessary (all models need to be considered), cf. Bell and Johnson-Laird (1998) and Evans, Handley, Harper, and Johnson-Laird (1999). Sixth, a mental model represents one possibility, and naive individuals tend to assume that each mental model is equiprobable unless there is evidence to the contrary (Johnson-Laird et al., 1999).

Some of these results can be accommodated by theories based on formal rules of inference. However, the most striking phenomenon predicted by the model theory is the existence of "illusory" inferences, and these errors cannot be explained by current formal rule theories. The illusions derive from the theory's fundamental representational principle, the principle of *truth*: individuals tend to minimise the load on working memory by representing explicitly only what is true, and not what is false, for assertions containing sentential connectives.

This principle concerns falsity, not negation, which is a purely syntactic notion. The principle postulates that people tend not to represent falsity, but it allows that they will represent negative assertions provided that they are true. The principle applies at two levels. At the first level, mental models represent only what is true, that is, each model represents a possibility given the premises. At the second level, however, mental models represent only those *literal* propositions in the premises that are true within each possibility. "Literal" is a technical term referring either to atomic sentences (those that contain no sentential connectives such as “and”, “if”, and “or”) or to their negations. Thus, 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 mental models representing two possibilities:

¬ 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 does not make explicit that it is false that there was an ace in this possibility; and the second model 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 forego the notation here, because reasoners rapidly forget these footnotes. There is a premium on truth in human reasoning.

If individuals represent what is true, do they never represent what is false? That cannot be the case. Individuals can construct models of the falsity of assertions that do not contain connectives. An assertion, such as:

The cup is on the right of the saucer

has a mental model that captures the spatial relation between the two objects, as in a plan:

saucer   cup

There are many ways in which the assertion could be false, e.g., if the cup were on the saucer or to the left of it. But the theory allows, as we have seen, that individuals can use negation as an operator on a model. Hence, they can represent the possibilities in which the assertion is false by constructing the mental model:

¬: saucer   cup

where the negation applies to the model as a whole. Individuals even have the option of imagining the different possibilities that would be instances of this negated model. That is, they can construct a model of the cup in the saucer, a model of the cup above the saucer, and so on. But they are only likely to do so when the falsity of an assertion is compatible with just a single possibility. For example, the falsity of a *negative* assertion, such as "The cup is not on the right of the saucer", is compatible with only the single possibility represented by the unnegated model given earlier. Hence, with true assertions, affirmatives are compatible with one possibility and negatives are compatible with many possibilities, whereas with false assertions, affirmatives are compatible with many possibilities and negatives are compatible with one possibility. This pattern may account for the well-established interaction in the verification of assertions: People are faster to evaluate true affirmatives than false affirmatives, but faster to evaluate false negatives than true negatives (Clark & Chase, 1972; Wason, 1959, 1961).

An important developmental milestone is the ability to pass the "false beliefs" test. Thus, 4-year-olds are usually able to distinguish their own beliefs from the false beliefs of another individual who failed to observe some key event (e.g., Leslie, 1994). Children are likewise able to engage in pretence, to detect lies and deception, and to reason on the basis of counterfactual suppositions. Ultimately, they may even be able to carry out Wason's selection task (e.g., Wason & Johnson-Laird, 1972). In all these cases, as Denise Cummins (personal communication) points out, reasoners need to represent what is false. However, a crucial feature of the false beliefs in studies of the "theory of mind" is that they do not depend on sentential connectives, i.e. they are simple "atomic" beliefs (Alan Leslie, personal communication).

Adults, of course, need to consider what is false in order to represent counterfactual conditionals, such as: "If there had been a king in the hand then there would have been an ace in the hand" (Byrne, 1997). They also need to consider what is false in order to search for counterexamples to conclusions (as Jonathan Evans, personal communication, has reminded us). But it is difficult for them to envisage the possibilities corresponding to the falsity of a compound assertion containing more than one sentential connective. They even make mistakes in listing the conditions in which conjunctions would be false. The evidence suggests that they do not have direct access to false possibilities but must infer them from true possibilities, and that they often err in the process (e.g., Barres & Johnson-Laird, 2000; Johnson-Laird & Barres, 1994).

The principle of truth predicts that certain premises will lead individuals to make "illusory inferences". The failure of mental models to represent what is false should lead to compelling, but fallacious, conclusions. Consider the following problem about a particular hand of cards:

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

Nearly everyone infers:

> There is an ace in the hand

and is highly confident that they are correct (Johnson-Laird & Savary, 1999). They think about the salient possibility in which the first conditional is true:

> King Ace

and they think about the salient possibility in which the second conditional is true:

> ¬ King Ace

It seems that there must be an ace in the hand, and this conclusion is reinforced by the categorical assertion that there is a king, which eliminates the second model. But, the conclusion is a fallacy granted a disjunction—exclusive or inclusive—between the two conditionals, and granted that when a conditional is false its antecedent can be true and its consequent false. In this case, one or other of the two conditionals may be false, and so there is no guarantee that there is an ace even though there is a king. If, for example, the first conditional is false, then there can be a king in the hand without an ace.

A single illusion, such as the preceding example, is open to several alternative explanations, particularly as the interpretation of conditionals is highly controversial. Over and Evans (1997), for example, argue that the inference is an illusion only granted an interpretation of conditionals as material implications. In our view, however, it is an illusion provided that the falsity of a conditional, such as:

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

yields no guarantee of an ace in the hand given that there is a king in the hand. Another possibility is that individuals overlook the exclusive disjunction or treat it as though it concerned only the antecedents of the conditionals (Lance Rips, personal communication). One aim of the studies in the present paper was accordingly to test illusions that depend, not on conditionals, but on disjunctions, which have a relatively uncontroversial meaning. Another aim was to examine a potential antidote to the illusions. A successful antidote is revealing because it illuminates the cause of the original illusion. The model theory predicts that the illusions arise from the failure to represent what is false, and so our antidote was designed to enable them to avoid this problem.

The form of the illusion that we investigated is illustrated by the following example:

1. *Illusion (logical disjunction)*

   Only one of the two following assertions is true about John:

   John is a lawyer or an economist, or both.
   John is a sociologist or an economist, or both.

   He is not both a lawyer and a sociologist.
   Is John an economist?

An assertion of the form "Only one of the two following assertions is true ..." makes a metalinguistic assertion about the truth of other assertions. In this case, it is logically equivalent to an exclusive disjunction between the two assertions, but it makes clearer that only one of them is true. The model theory predicts, however, that naive individuals—that is, those with no training in logic—will build models that represent only what is true. They will accordingly envisage the following possibilities from first the disjunction:

lawyer	
	economist
lawyer	economist

and the following possibilities from the second disjunction:

sociologist	
	economist
sociologist	economist

The second premise asserts that John is not both a lawyer and a sociologist, and so it is consistent with these models. Naive individuals leave open the possibility that John is, or is not, an economist. The model theory accordingly predicts that reasoners should tend to infer that one cannot tell whether John is an economist. Naive reasoners may be confused by negated conjunctions of the form, "... is not both A and B", and treat them as equivalent to conjunctions of negations: "not A and not B" (Barres & Johnson-Laird, 2000). In this case, they will infer that John is not a lawyer and not a sociologist. When these cases are eliminated from the models just given, only one possibility remains:

economist

and so such reasoners will respond, "Yes" (John *is* an economist).

Both the "Yes" response and the "cannot tell" response are illusions. When the first disjunction is true, the second disjunction is false, and so there is the following possibility for John:

lawyer	¬ economist	¬ sociologist

And when the second disjunction is true, the first disjunction is false, and so there is the following possibility for John:

¬ lawyer	¬ economist	sociologist

It follows that John is not an economist, and the second premise that John is not both a lawyer and a sociologist adds no new information. The advantage of this sort of illusory inference is that the meaning of disjunctions, unlike those of conditionals, is straightforward.

The experiment also examined a control inference for which the failure to represent what is false should have no effect on correct performance. This inference was based on the same initial disjunction, but the second premise was a categorical conjunction, and the question concerned a different predicate. This inference is illustrated by the following example:

2. *Control (logical disjunction)*
   Only one of the two following assertions is true about John:
   John is a lawyer or an economist, or both.
   John is a sociologist or an economist, or both.
   He is not a lawyer and he is not an economist.
   Is John a sociologist?

Given the initial mental models of the two disjunctions (see earlier):

lawyer	
	economist
lawyer	economist

and:

sociologist	
	economist
sociologist	economist

the second premise eliminates every case except the following:

sociologist

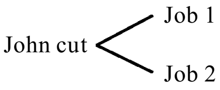
and so the model theory predicts that naive reasoners will respond, "Yes" (John is a sociologist). The fully explicit models of the premises, which take falsity into account, support the same conclusion, and so it is valid.

The model theory predicts that the illusions will occur because naive reasoners represent what is true, but not what is false. If so, what is likely to help reasoners to avoid the illusions? One potential antidote is to provide a physical cue in which the disjunction is between two separate entities. In this way, when one entity is chosen, it should be clear that the other entity has not been chosen. An analogous phenomenon occurs in reasoning about whether a protagonist has uttered the truth or a falsehood. It is easier to reason about such problems when the protagonist refers to physical entities than to the truth status of their own or other's assertions (Byrne, Handley, & Johnson-Laird, 1995). In the present case, we can re-express both the preceding problems (the illusory and the control inference) so that they concern physical objects, by referring to advertisements in a newspaper. Here is the "physical" version of the illusory inference:

3. *Illusion (physical disjunction)*

   John was reading the newspaper looking for a job. There were two ads on that page but he cut out only the one that matched his qualifications:
   Job 1 was for a lawyer or an economist, or both.
   Job 2 was for a sociologist or an economist, or both.
   He is not both a lawyer and a sociologist.
   Is John an economist?

It should be easy to appreciate that if John cuts out the advertisement for Job 1, then he did not cut out the advertisement for Job 2, and vice versa. The mental models of the rubric are, in effect:



This notation represents that John cut out either the advertisement for Job 1 or the advertisement for Job 2. And this disjunction between entities should be easier to grasp than the disjunction between propositions as a whole in the original illusory inference, that is, if the first disjunction is true, then the second disjunction is not true, and vice versa. The control problem for the physical entities is illustrated by the following example:

4. *Control (physical disjunction)*

   John was reading the newspaper looking for a job. There were two ads on that page but he cut out only the one that matched his qualifications:
   Job 1 was for a lawyer or an economist, or both.
   Job 2 was for a sociologist or an economist, or both.
   He is not a lawyer and he is not an economist.
   Is John a sociologist?

In summary, Experiment 1 was designed to test two predictions: first, illusory inferences should occur with disjunctions of assertions about an individual; and, second, that the illusions should be reduced when the exclusive disjunction (between the two disjunctive premises) concerns physical entities rather than truth values.

EXPERIMENT 1

Method

Design. The participants acted as their own controls and carried out four main sorts of inference and three filler items. The filler items on the second, fourth, and sixth trials were easy problems to which the correct answer was "no". The four inferences were presented to each participant on the remaining trials in a counterbalanced order determined by a Williams' square to control for order and residual effects. The four inferences were based on an illusory and a control inference presented either in the form of a logical disjunction or a disjunction of physical entities. The illusory inferences had the following logical form:

Only one of the following is true/Only one of the following entities was selected:
  Ea or Fa, or both.
  Ga or Fa, or both.
Not both Ea and Ga.
Is Fa the case?

where "a" is a proper name, and "E", "F", and "G" are predicates. The control inferences had the following form:

Only one of the following is true/Only one of the following entities was selected:
  Ea or Fa, or both,
  Ga or Fa, or both.
Not-Ea and not-Fa.
  Is Ga the case?

The four inferences are illustrated in the Introduction.

Materials and procedure. The present experiment and the subsequent experiments were carried out in Spanish with native speakers of the language. We devised a set of 16 problem contents, which all concerned a named protagonist and three common vocations. They were assigned twice at random to the four main problems, and half the participants tested with one set of materials and the other half tested with the other set of materials. No individual encountered a particular content more than once in the experiment.

Each participant received a booklet of the seven problems, one problem on each page. The first page contained the instructions. They stated that the participants' task was to read some information about an individual and then to answer a question, such as: "Is John a sociologist?". They were to respond "yes" if they thought that the conclusion must be true given the premises, "no" if they thought that the conclusion must be false given the premises, and "cannot tell" if they thought that the premises did not provide enough information to decide between the two possibilities. The instructions explained that the first premise had the rubric that only one of the following assertions was true, that is, one assertion was true and the other was not true, and the participants were told that it was important to bear this point in mind. After the participants had made their response, they rated their confidence in their answers on a 5-point scale, ranging from 1 ("no confidence") to 5 ("full confidence").

Participants. We tested 32 undergraduate students in psychology from the University of La Laguna, who received course credit for their participation. They had no training in logic and had not participated before in any experiment on deductive reasoning.

Results and discussion

Table 1 presents the percentages of correct responses in each condition, together with the participants' mean confidence ratings. The control inferences (98% correct) were reliably easier than the illusory inferences (16% correct; Wilcoxon test, $z = 5.07$; $p < .0001$, one-tailed). The physical disjunctions (61% correct) were reliably easier than the logical disjunctions (53% correct, Wilcoxon test, $z = 1.89$; $p < .03$, one-tailed). And the interaction between the two variables was significant (Wilcoxon test, $z = 2.65$; $p < .005$, one-tailed). As predicted, the interaction arose because the illusory inferences were reliably easier with physical disjunctions than with logical disjunctions (Wilcoxon, $z = 2.45$, $p < .007$, one-tailed), whereas the control inferences were not reliably easier with physical disjunctions than with logical disjunctions. The logical disjunctions appeared to yield more compelling illusions because 31% of responses were "yes", whereas only 19% of responses to the physical disjunctions were "yes". The balance of the responses were "cannot tell" (56% with the physical disjunctions and 63% with the logical disjunctions). The percentage of correct conclusions for the filler problems was 95%. There was no reliable sign of any transfer from the physical disjunctions to the logical disjunctions.

TABLE 1
Experiment 1

	Illusory inferences	*Control inferences*
Logical disjunctions	6 (4.3)	100 (4.8)
Physical disjunctions	25 (4.4)	97 (4.9)

The percentages of correct conclusions in Experiment 1 (n = 32), and the mean confidence ratings (1 = "no confidence" and 5 = "full confidence").

The participants were highly confident in their responses (an overall mean of 4.5 on a 5-point scale). They were more confident in their responses to the control problems (4.9) than in their responses to the illusions (4.4: Wilcoxon test, $z = 3.57$; $p < .0001$, two-tailed). They were slightly more confident in their responses to the physical disjunctions (4.7) than in their responses to the logical disjunctions (4.5), but the difference was only marginally reliable (Wilcoxon test, $z = 1.78$, $p = .075$, two-tailed). The interaction between these two variables was not significant (Wilcoxon, $z = .24$). We computed the Spearman rank-order correlation between the correctness of the response (whether or not it was correct) and the confidence ratings for the illusory inferences: 0.18 for the physical disjunctions and –0.07 for the logical disjunctions. Neither correlation was significant. The correlations could not be computed for the control inferences, because there were too few incorrect responses.

The main result is that illusory inferences occurred with the present disjunctive problems. Unlike conditionals, the semantics of disjunctions is straightforward—a claim that applies both to the exclusive disjunction expressed by the rubrics "Only one of the following is true" and "Only one of the following entities was selected", and to the individual disjunctions of the form "John is a lawyer or an economist, or both". The illusions are arguably slightly more complex logically because they contain a premise of the form:

Not both p and q

whereas the corresponding premise in the control problems was:

Not-p and not-q.

However, the negated conjunction cannot be responsible for the illusions, because a proper interpretation of the disjunctive premise alone yields the correct response. In contrast, the participants who followed the principle of truth will either respond "nothing follows" (63% of trials) if they interpreted the negated conjunction correctly, or respond "yes" (31% of trials) if they interpreted it incorrectly as: not-p and not-q. (Experiment 3 examines simple conjunctions in both illusory and control problems.) Moreover, the negated conjunction is common to both physical and logical disjunctions, and a second result is that the participants were less likely to err when the premises referred to a physical disjunction between two advertisements than to a logical disjunction between the truth values of assertions. The effect was not large (25% correct versus 6% correct), but it was reliable. Most participants erred by choosing the "cannot tell" option, and this choice might reflect a certain wariness on their part, which also showed up in the reliably lower confidence ratings. In order to examine this possibility, we excluded the response option of "cannot tell" from the next experiment.

EXPERIMENT 2

Method

This experiment replicated the first experiment, but excluded the response option "cannot tell". The model theory predicts that the participants should still succumb to the illusions. Performance with the physical disjunctions, however, should be more accurate than with the logical disjunctions. The design and materials were identical to Experiment 1. The procedure was the same except that there were only two response options, "yes" (the conclusion follows from the premises) and "no" (the conclusion does not follow from the premises).

Participants. We tested a new sample of 32 students from the same population as before.

Results and discussion

Table 2 presents the percentages of correct responses in each condition, together with the participants' mean confidence ratings. Once again, the control inferences (97% correct) were reliably easier than the illusory inferences (30% correct; Wilcoxon test, z = 2.92; p < .002, one-tailed). However, in this experiment, there was no difference in accuracy between the physical and the logical disjunctions (Wilcoxon test, z = 0.58), nor was the interaction between the two variables significant (Wilcoxon, z = 0.58). The pattern in the confidence ratings was the same. The participants were more confident in their responses to the control problems (4.7) than in their responses to the illusions (4.1; Wilcoxon test, z = 3.12; p < .002, two-tailed), and no other difference was reliable. The Spearman correlation between confidence and performance was 0.02 for the illusory inferences with physical disjunctions, it was 0.24 for illusory inferences with logical disjunction experimental condition; and neither correlation was statistically significant. The percentage of correct conclusions for the filler problems was 98%.

The results confirmed that naive reasoners succumb to the illusory inferences, inferring that a particular conclusion holds when, in fact, it is invalid. They made

TABLE 2
Experiment 2

	Illusory inferences	*Control inferences*
Logical disjunctions	31 (4.1)	97 (4.7)
Physical disjunctions	28 (4.1)	97 (4.6)

The percentages of correct conclusions in Experiment 2 (n = 32), and the mean confidence ratings (1 = "no confidence" and 5 = "full confidence").

such errors even when they were no longer offered the option of responding "cannot tell". However, an unexpected effect of dropping this option was to eliminate any advantage of physical disjunctions over logical disjunctions. What happened was that performance with the logical disjunctions appeared to improve up to the level of the physical disjunctions. It is dangerous to compare the results of different experiments, but one potential explanation (which we owe to David Green) is that some participants guess—at least with the illusory problems—and so they are liable to do better when there are only two response alternatives. This failure of the antidote to produce an enhanced effect on the physical disjunctions was contrary to our original hypothesis, and forced us to re-examine Experiment 1, which yielded only a small advantage for them. One reason may have been that although the first premise in the physical disjunctions referred to separate entities (the advertisement that the protagonist cut out), neither the second premise nor the question did. They merely referred to the protagonist, e.g.:

He is not both a lawyer and a sociologist.
John an economist?

Our next experiment accordingly modified the problems based on physical disjunctions so that each premise and the question referred to the physically separate entities.

EXPERIMENT 3

Method

The previous experiments corroborated the occurrence of illusions, and suggested that physical disjunctions can sometimes work as an antidote. They yielded a small but reliable effect in Experiment 1. As we argued earlier, one reason for the small effect might have been that the physical disjunctions occurred only in the first premise. In Experiment 3, we therefore ensured that the physical disjunctions referred to separate entities in all the premises and the conclusion. In Experiments 1 and 2, there was a difference in the logic of the conjunctive premises between the illusions and control problems. In the design of Experiment 3, we therefore used conjunctions of the same form for both the illusions and the controls. Experiment 2 used only two response options, which improved performance with the logical illusions, but not the physical illusions. One reason, as we suggested, might have been the consequences of guessing. We therefore retained two response options in the present experiment, but sought to make the task subjectively easier in order to reduce guessing (with an expected reduction in the difference between the confidence ratings of responses to illusions and controls). We aimed to do so by using conclusions that concerned only possibilities (Bell & Johnson-Laird, 1998; Goldvarg & Johnson-Laird, 2000). In order to use such conclusions in both illusions and controls, it was necessary to use disjunctions of three assertions.

In summary, we made three changes to the problems in order to examine the illusions and their potential antidote in a more uniform way. First, the problems based on physical disjunctions referred to separate entities in all the premises and the conclusion. Second, the conjunctions were the same for both illusions and controls. And third, the main disjunction concerned three assertions and the conclusion concerned something that was possible. The four sorts of problem are illustrated by the following examples:

1. *Illusion (logical disjunction)*

Only one of the three following assertions is true about John:
1. John is a physicist or an engineer, or both.
2. John is an architect or an engineer, or both.
3. John is a biologist.

John is not a physicist and he is not an architect.
Consequently, is it possible that John is an engineer?

The disjunction of the three assertions has the following mental models:

physicist	
	engineer
physicist	engineer

or:

architect	
	engineer
architect	engineer

or:

biologist

The categorical assertion eliminates all the models apart from:

engineer
biologist

Hence, the model theory predicts that naive reasoners should respond "yes" to the question. The response is an illusion. If John were an engineer then two of the premises in the initial disjunction would be true (1. and 2.) contrary to the rubric that only one of them is true.

2. *Control (logical disjunction)*

   Only one of the three following assertions is true about John:
   1. John is a physicist or an engineer, or both.
   2. John is an architect or an engineer, or both.
   3. John is a biologist.

   John is not a physicist and he is not a biologist.
   Consequently, is it possible that John is an architect?

This problem has the same disjunctive premise (and therefore mental models) as the illusory inference, but the categorical assertion has a different second constituent proposition. It rules out all the mental models except the following ones:

architect	
	engineer
architect	engineer

The model theory therefore predicts that naive reasoners should respond "yes" to the question. This response is supported by the one fully explicit model of the premises:

architect	¬ physicist	¬ engineer	¬ biologist

and so the response is correct.

3. *Illusion (physical disjunction)*

   John was reading the newspaper looking for a job. There were three ads on that page, but John cut out only the one that matched his profession:

   Ad 1 asked for a physicist or an engineer, or both.
   Ad 2 asked for an architect or an engineer, or both.
   Ad 3 asked for a biologist.

   John didn't cut out an ad asking for a physicist and he didn't cut out an ad asking for an architect.
   Consequently, is it possible that John cut out an ad that asked for an engineer?

The problem is equivalent to the illusion based on a logical disjunction, but as it concerns a disjunction of separate physical entities, naive reasoners should tend to succumb to the illusory "yes" response less often than in the case of the logical disjunction.

4. *Control (physical disjunction)*

   John was reading the newspaper looking for a job. There were three ads on that page, but John cut out only the one that matched his profession:

> Ad 1 asked for a physicist or an engineer, or both.
> Ad 2 asked for an architect or an engineer, or both.
> Ad 3 asked for a biologist.
>
> John didn't cut out an ad asking for a physicist and he didn't cut out an ad asking for a biologist.
>
> Consequently, is it possible that John cut out an ad that asked for an architect?

The problem is equivalent to the control problem based on a logical disjunction, and so reasoners should tend to respond "yes" correctly. In summary, the control problems should yield a greater percentage of correct responses, but the physical disjunctions should act as an antidote to the illusory inferences. The design and contents of the materials were identical to those of the previous experiments. The procedure was the same as Experiment 2, i.e., there were only two response options; "yes" (the conclusion follows from the premises) and "no" (the conclusion does not follow from the premises).

Participants. We tested a new sample of 32 students from the same population as before.

Results

Table 3 presents the percentages of correct responses in each condition, together with the participants' mean confidence ratings. The control inferences (100% correct) were reliably easier than the illusory inferences (52% correct; Wilcoxon test, $z = 4.77$, $p < .0001$, one-tailed). The physical disjunctions (92% correct) were reliably easier than the logical disjunctions (59% correct; Wilcoxon test, $z = 4.58$; $p < .0001$, one-tailed). And the interaction between the two variables was significant (Wilcoxon test, $z = 4.58$; $p < .0001$, one-tailed). As predicted, the interaction arose because the illusory inferences were reliably easier with physical disjunctions than with logical disjunctions (Wilcoxon, $z = 4.58$; $p < .0001$, one-tailed), whereas the control inferences were not reliably easier with physical disjunctions than with logical disjunctions. The percentage of correct conclusions for the filler problems was 95%.

TABLE 3
Experiment 3

	Illusory inferences	*Control inferences*
Logical disjunctions	19 (4.5)	100 (4.7)
Physical disjunctions	84 (4.9)	100 (4.9)

The percentages of correct conclusions in Experiment 3 (n = 32), and the mean confidence ratings (1 = "no confidence" and 5 = "full confidence").

The participants were highly confident in their responses (an overall mean of 4.7 on a 5-point scale). Their confidence in their responses to the control problems (4.8) did not differ reliably from their confidence in their responses to the illusions (4.8: Wilcoxon test, $z = 1.37$). However, they were slightly more confident in their responses to the physical disjunctions (4.9) than in their responses to the logical disjunctions (4.6; Wilcoxon test, $z = 2.5$; $p < .02$, two-tailed). The interaction between these two variables was not significant (Wilcoxon, $z = 1.1$). We computed the Spearman rank-order correlation between the correctness of the response (whether or not it was correct) and the confidence ratings for the illusory inferences: 0.16 for the physical disjunctions and 0.05 for the logical disjunctions. Neither correlation was significant. The correlations could not be computed for the control inferences, because there were too few incorrect responses. As in the previous experiments, there was no reliable sign of any transfer from the physical disjunctions to the logical disjunctions.

The experiment confirmed the occurrence of illusions. It also showed that they are reliably reduced when the disjunctions occur over physical objects rather than assertions. Performance in this condition was well above chance. Also, the participants were highly confident in all their responses, including their erroneous ones with the logical illusions. The experiment therefore established an effective antidote to illusory inferences.

GENERAL DISCUSSION

The experiments established that naive reasoners are susceptible to illusory inferences based on disjunctions of disjunctions (cf. Goldvarg & Johnson-Laird, 2000). Experiment 1 showed that given a problem of the form:

> Only one of the two following assertions is true about John:
> John is a lawyer or an economist, or both.
> John is a sociologist or an economist, or both.
> He is not both a lawyer and a sociologist.
> Is John an economist?

the majority of participants wrongly inferred either that the answer was “yes” or that “one cannot tell”. In fact, John cannot be an economist because the first premise would be false in that case. Nearly 70% of participants responded “yes” in Experiment 2 when the “cannot tell” option was removed. Experiment 3 corroborated these results, showing that illusory inferences occur with problems of the form:

> Only one of the three following assertions is true about John:
> 1. John is a physicist or an engineer, or both.
> 2. John is an architect or an engineer, or both.
> 3. John is a biologist.

John is not a physicist and he is not an architect.
Consequently, is it possible that John is an engineer?

Again, the majority of participants responded "yes", although if John were an engineer then the disjunctive premise would be false.

In Experiments 1 and 2, the participants were slightly less confident in their illusory conclusions than in their correct control responses. We now suspect that their wariness arose because the premise:

John is not both a lawyer and a sociologist

may be slightly harder to understand (it calls for models of three possibilities) than the corresponding premise in the control problem:

John is not a lawyer and he is not an economist

which calls for only a single model. This hypothesis is borne out by the results of Experiment 3. Both the illusory and the control problems had a second premise that was a simple conjunction (as in the preceding example), and the participants did not differ reliably in their confidence ratings to the two sorts of problem.

The second conclusion supported by the experiments is that there is an effective antidote to the disjunctive illusions. If someone tells you that you may have either soup or else salad, but not both, you should be more likely to bear in mind that if you take the salad then you forego the soup. In other words, it is easier to cope with a disjunction of physical objects than with a disjunction of truth values of assertions. The effect is not merely that one is dealing with concrete imageable objects, because illusions occur with premises that refer, for example, to playing-cards, which are easy to imagine. It is a result of a disjunction of physical entities rather than a disjunction of propositions. This intuition lay behind the contrast between the physical and logical disjunctions that we manipulated in the three experiments. Experiment 1 demonstrated a small but reliable improvement in performance with the physical disjunctions, but it was not enhanced when we removed the "cannot tell" option in Experiment 2. We then modified the problems so that all the premises in the physical disjunction problems referred explicitly to the relevant entities. The result was a large improvement in performance with the illusions from 19% correct with logical disjunctions to 84% correct with physical disjunctions.

If our account is correct then illusory inferences are not merely a feature of abstract thought. They arise from the principle of truth, which postulates that people tend to think about what is true rather than what is false. Hence, an exclusive disjunction of *propositions* that concern physical objects should give rise to illusions, and evidence corroborates this prediction (Goldvarg &

Johnson-Laird, 2000). Likewise, although disjunctions of physical objects reduce the illusions, other ways to reduce them also exist. Yang and Johnson-Laird (in press), for example, reduced illusions in reasoning with quantifiers by teaching the participants to work through the consequences of one assertion being true and another assertion being false. The drawback with this procedure was that, not surprisingly, it made the control problems harder.

Is there any other explanation for the illusions, apart from the failure to represent falsity? Sceptics may suspect that participants overlook the rubric, “Only one of the following assertions is true”, and accordingly treat the disjunctions as though they are in a conjunction. We took pains, however, to emphasise in the instructions that one and only one of the disjunctions was true. Indeed, a large proportion of participants in Experiment 1 chose the erroneous “cannot tell” option, which is incompatible with a conjunctive interpretation. The success of the antidote counts against the interpretation too. Moreover, our colleague Bonnie Meyer has run an unpublished study using the same rubric but in which the participants have to think aloud. It was obvious from their protocols that they grasped the disjunctive nature of the rubric. We conclude that the most likely cause of the illusions is the failure to represent what is false.

Current formal rule theories neither predict the illusory inferences nor accommodate them *post hoc*. These theories rely solely on valid rules of inference (e.g., Braine & O’Brien, 1998; Rips, 1994), 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 strategy for making suppositions, i.e. assumptions made for the sake of argument (Luca Bonatti and David O’Brien, personal communications). This hypothesis has some plausibility for certain illusions, but it makes the wrong predictions for other inferences (Johnson-Laird, Legrenzi, Girotto, & Legrenzi, 2000; Yang & Johnson-Laird, 2000, in press). It is also unable to account for the present illusions based on disjunctions.

Previous studies of illusory inferences have sometimes concerned conditional premises (Johnson-Laird & Savary, 1999). They yield powerful effects, and such illusions even occur in daily life. A search of the worldwide web turned up several examples, including the case of a professor who warned students who were absent from class:

> … either a grade of zero will be recorded if your absence is not excused, or else if your absence is excused other work you do in the course will count.

This assertion has the following mental models:

¬ excused	zero-grade	
excused		other-work

and presumably both the professor and the students had these two possibilities in mind. Yet they are illusory. What the professor should have asserted was "*but* if your absence …" which has the force of a conjunction, i.e. both conditionals are true. The use of "or else" means that one of the conditionals *as a whole* is true and one of them is false, which has quite unintended consequences, e.g., students may not get a zero grade even when they have no excuse for their absence. Errors of this sort seem quite pervasive. If readers have every wondered whether to use "or" or "and" in a sentence, then they too may have been tempted by an illusory inference. We have observed spontaneous illusory inferences drawn by individuals as they reason aloud about problems based on sentential connectives (Van der Henst, Yang, & Johnson-Laird, 2000). The failure to consider falsity is likely to have more serious consequences in a technological setting.

From a psychological point of view, illusory inferences based on conditionals are mildly problematic, because of the lack of consensus about the meaning of conditionals. In contrast, the semantics of disjunctions are straightforward. "Only one of the … following assertions is true" is equivalent to an exclusive disjunction, which is false if more than one of the assertions is true. Likewise, an assertion of the form: "John is a lawyer or an economist, or both" is clearly true in three cases: John is a lawyer and not an economist, John is an economist and not a lawyer, and John is both a lawyer and an economist. It follows that in an exclusive disjunction of such disjunctions, a protagonist cannot satisfy any predicate that occurs in more than one disjunction. Yet as our results show, naive reasoners fail to make this inference. When they imagine one disjunction as true, they do not represent explicitly the falsity of any other disjunction.

Manuscript received 23 November 1999
Revised manuscript received 2 June 2000

REFERENCES

Barres, P.E., & Johnson-Laird, P.N. (2000). *On imagining what is true (and what is false)*. Manuscript submitted for publication.

Bauer, M.I., & Johnson-Laird, P.N. (1993). How diagrams can improve reasoning. *Psychological* Science, *4*, 372–378.

Bell, V., & Johnson-Laird, P.N. (1998). A model theory of modal reasoning. *Cognitive Science*, *22*, 25–51.

Braine, M.D.S., & O'Brien, D.P. (Eds.) (1998). *Mental logic.* Mahwah, NJ: Lawrence Erlbaum Associates Inc.

Bucciarelli, M., & Johnson-Laird, P.N. (1999). Strategies in syllogistic reasoning. *Cognitive Science*, *23*, 247–303.

Byrne, R.M.J. (1997). Cognitive processes in counterfactual thinking about what might have been. In D.K. Medin (Ed.) *The psychology of learning and motivation. Advances in research and theory* (Vol. 37, pp. 105–154). San Diego, CA: Academic Press.

Byrne, R.M.J, Handley, S.J., & Johnson-Laird, P.N. (1995). Reasoning from suppositions. *Quarterly Journal of Experimental Psychology*, *48*A, 915–944.

Clark, H.H., & Chase, W.G. (1972). On the process of comparing sentences against pictures. *Cognitive Psychology*, *3*, 472–517.

Evans, J.St.B.T., Handley, S.J., Harper, C.N.J., & Johnson-Laird, P.N. (1999). Reasoning about necessity and possibility: A test of the mental model theory of deduction. *Journal of Experimental Psychology*: *Learning, Memory, and Cognition*, *25*, 1495–1513.

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

Goldvarg, Y., & Johnson-Laird, P.N. (2000). Illusions in modal reasoning. *Memory & Cognition*, *28*, 282–294.

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*, 475–478.

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

Johnson-Laird, P.N. Legrenzi, P., Girotto, V., & Legrenzi, M.S. (2000). Illusions in reasoning about consistency. *Science*, *288*, 531–532.

Johnson-Laird, P.N., Legrenzi, P., Girotto, V., Legrenzi, M., & Caverni, J-P. (1999). Naive probability: A mental model theory of extensional reasoning. *Psychological Review*, *106*, 62–88.

Johnson-Laird, P.N., & Savary, F. (1999). Illusory inferences: A novel class of erroneous deductions. *Cognition*, *71*, 191–229.

Leslie, A.M. (1994). Pretending and believing: Issues in the theory of ToMM [Theory of Mind Mechanism]. *Cognition*, *50*, 211–238.

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

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

Schaeken, W.S., & Johnson-Laird, P.N. (2000). *Temporal reasoning and the size of mental models*. Manuscript submitted for publication.

Van der Henst, J-B., & Yang, Y., & Johnson-Laird, P.N, (2000). *Strategies in sentential reasoning*. Manuscript submitted for publication.

Wason, P.C. (1959). The processing of positive and negative information. *Quarterly Journal of Experimental Psychology*, *21*, 92–107.

Wason, P.C. (1961). Response to affirmative and negative binary statements. *British Journal of Psychology*, *52*, 133–142.

Wason, P.C., & Johnson-Laird, P.N. (1972). *Psychology of reasoning: Structure and Content*. Cambridge, MA: Harvard University Press.

Yang, Y., & Johnson-Laird, P.N. (in press). Systematic fallacies in quantified reasoning and how to eliminate them. *Memory & Cognition*.

Yang, Y., & Johnson-Laird, P.N. (2000). Illusions in quantified reasoning: How to make the impossible seem possible, and *vice versa*. *Memory & Cognition*, *28*, 452–465.