

Mental models, sentential reasoning, and illusory inferences

P.N. Johnson-Laird¹

Department of Psychology, Princeton University²

Abstract

This chapter describes how individuals reason with sentential connectives, such as “if”, “or”, and “and”. They do not have a “truth functional” semantics for these connectives, but rather they construct models of the possibilities compatible with sentences in which the connectives occur. Human working memory has a limited processing capacity, and so individuals aim to construct only a single model at a time, and to represent only those clauses in the premises that hold in each possibility. One unexpected consequence of the theory emerged from its computer implementation. Certain inferences should yield systematic fallacies if reasoners use mental models. The chapter explains this prediction and reports some studies corroborating the occurrence of these “illusory” inferences. No one has yet devised an account of them on the basis of another theory.

Suppose that you are carrying out a test of system and you know that if the test is to continue then the reactivity of the system must not have reached the critical level. You then observe that the reactivity has reached the critical level. What should you do? It seems obvious that you should stop the test. The engineers in charge at Chernobyl were in this position, but they continued the test (see Medvedev 1990). Why they continued is puzzling, because the test was not only dangerous, but pointless. It led to

¹ This research was supported by a grant from the National Science Foundation (BCS-0076287) to study strategies in reasoning. For their helpful advice, I thank Ruth Byrne, Vittorio Girotto, Geoff Goodwin, Uri Hasson, Karl Christoph Klauer, Louis Lee, Markus Knauff, Walter Schroyens, André Vandierendonck, Clare Walsh, and Yingrui Yang.

² E-mail: phil@princeton.edu

the disaster. One possibility is that the engineers failed to make a valid inference of the form:

If A then not B.

B.

Therefore, not A.

where *A* stands for “the test is to continue” and *B* stands for “the reactivity has reached the critical level”.

For several years, I have given groups of engineering students a similar problem with an abstract content, such as:

If there is a triangle on the board then there is a circle on the board.

There isn’t a circle on the board.

What, if anything, follows?

Typically, more than half of them respond that nothing follows from these premises. In fact, the premises yield the conclusion:

There is not a triangle on the board.

This conclusion is *valid*: it must be true given that the premises are true. But, the inference is quite difficult to make. The engineers are not reluctant to make inferences, because with premises of this sort:

If there is a triangle on the board then there is a circle on the board.

There is a triangle on the board.

nearly all of them draw the valid conclusion:

There is a circle on the board.

People do make mistakes, and the difference in difficulty between the two previous inferences is one of the most robust effects in the psychology of reasoning (see, e.g. Evans et al. 1993). Yet, reasoners are not always wrong. Psychologists therefore need to explain both their logical ability and the cause of their mistakes.

My aim in this chapter is to describe the mental mechanisms underlying a major sort of reasoning, so-called “sentential reasoning”, which is based on negation and sentential connectives, such as “if”, “or”, and “and”. The account is a development from the theory of mental models (Johnson-Laird 1983, Johnson-Laird & Byrne 1991). The theory postulates that the mind constructs models of the world that it uses to reason. It constructs them from perception (Marr 1982), imagination (Metzler & Shepard 1982), knowledge (Gentner & Stevens 1983), and the comprehension of discourse (Stevenson 1993, Polk & Newell 1995, Oakhill & Garnham 1996, Garnham 2001). A crucial distinction between models and other sorts of proposed mental representation is that the structure of models corresponds to the structure of what they represent: individuals are represented by individual tokens, properties by properties of these tokens, and relations by relations among these tokens (see, e.g. Johnson-Laird 1983).

In reasoning, a key step is to establish a conclusion; its strength depends on whether any models of the premises refute it (Johnson-Laird & Byrne

1991). The theory therefore provides a unified account of reasoning about what is necessary, probable, or possible. A conclusion is *necessary* if it holds in all the models of the premises, it is *probable* if it holds in most models of the premises (Johnson-Laird et al. 1999), and it is *possible* if it holds in at least one model of the premises (Bell & Johnson-Laird 1998).

The model theory, as I refer to it, is based on a core principle that concerns the interpretation of connectives, and that gives rise to systematic fallacies. These fallacies can be so compelling that they have an illusory quality: it is hard to avoid succumbing to them even when you are on guard against them. You will understand the principle more easily if I outline elementary logic. Hence, the chapter begins with such an account. It then describes the interpretation of connectives in natural language, and illustrates the limitations of human working memory. These limitations lead to the fundamental principle of the model theory: mental models are parsimonious. The chapter formulates the mechanisms that implement this principle in the construction of mental models, which it contrasts with the reasoning of superhuman entities with unlimited working memories. It reports some illustrative results of recent studies of the illusory inferences. These results corroborate the theory.

1. Logic and truth-functional connectives

Logic treats sentences as expressing propositions; in everyday life, however, the proposition that a sentence expresses almost always depends on its context. “I can hear you now”—an utterance all too common these days—expresses different propositions depending on who says it, to whom it is addressed, and the time and circumstances of the utterance. To keep matters simple, I will use sentences that depend as little as possible on their context, and, where feasible, I will adopt the fiction that sentences are propositions.

Logic is the science of valid inferences. It is not concerned with how people make such inferences. Logicians have formulated many different calculi for formalized languages. They can set up a calculus in two distinct ways (see, e.g. Jeffrey 1981). The first way is formal, concerning patterns of symbols, but not their interpretation. The sentential calculus concerns sentential connectives in their logical senses—a notion that I will explain soon. Its *formal* specification depends on rules of inference, such as:

A or B, but not both.

not-B

Therefore, A.

Table 1

The truth table for an inclusive disjunction

<i>There is a circle on the board.</i>	<i>There is a triangle.</i>	<i>There is circle on the board or else a triangle or both.</i>
True	True	True
True	False	True
False	True	True
False	False	False

The variables, A and B , can have as values any declarative sentences whatsoever.

The second way to characterize a calculus is *semantic*. Consider an atomic sentence, i.e., one that contains neither negations nor connectives:

There is a circle on the board.

Let's suppose that it is false. A *compound* sentence is made from atoms by combining them with negation or sentential connectives. Here is a negative compound:

There is not a circle on the board.

This assertion is true because, as I just told you, the atom that it contains is false. Suppose that you also know another compound assertion, which is a disjunction of two atoms:

There is a triangle on the board or there is a circle, or both.

This disjunction is *inclusive*, because it allows that both atoms could be true. Hence, its meaning is compatible with three possibilities:

There is a triangle on the board and there is not a circle.

There is not a triangle on the board and there is a circle.

There is a triangle on the board and there is a circle.

You already know that there is not a circle, and so you can eliminate all but the first possibility. It follows that there is a triangle. The formal rule above also allows you to make this inference, but here you have made it on a semantic basis. Hence, in principle, human reasoning could be based on formal procedures or semantic procedures, or both.

The meaning of the preceding disjunction can be laid out in the form of a truth table, which specifies the truth value of the disjunction for each of the four possible contingencies—the three possibilities in which it is true, and the remaining possibility in which it is false. Table 1 presents this truth table. Each row in the table shows a possible combination of the truth values of the two atoms, and the resulting truth value of their inclusive disjunction. For example, the first row is the possibility in which both atoms are true, and, as a result, the inclusive disjunction is true too. Truth tables

were invented by the great American logician, Charles Sanders Peirce (see, e.g. Berry 1952), though Wittgenstein (1922) is often wrongly credited with their invention.

A sentential connective has a “logical” meaning when its interpretation can be summarized in a truth table. The truth table shows how the truth value of a sentence containing the connective depends solely on the truth values of its constituent propositions. Once you know their truth values, you can work out the truth value of the sentence as a whole from the connective’s truth table. Hence, an inclusive disjunction in its logical sense is true or false solely as a *function of* the truth values of the constituent propositions. As logicians say, a disjunction has a *truth-functional* meaning. This piece of jargon means: you feed in truth values of the constituent propositions, and the truth table for “or” gives an output of a truth value.

In logic, a general recipe exists for interpreting compound sentences. You replace each atom with its truth value—how you obtain such truth values is not part of the theory—and you progressively simplify the compound according to the interpretation of each connective, until you arrive at a final truth value for the sentence as a whole. This truth value depends only on the truth values of the atoms and the truth-functional interpretations of the connectives.

Here is an example of such an interpretation. Consider the compound assertion in which “or else” is an *exclusive* disjunction, i.e., only one of the two clauses it connects is true:

(A and not B) or else (C and D)

and assume that all the atoms are true: *A*, *B*, *C*, and *D*, are all true. The first step in the interpretation of the compound is to replace its atoms with their truth values:

(true and not true) or else (true and true)

The next steps simplify the expression according to the truth-functional meanings of negation and the connectives:

(true and false) or else (true and true)	—according to the meaning of <i>not</i>
(false or else true)	—according to the meaning of <i>and</i>
true	—according to the meaning of or <i>else</i>

Hence, the compound assertion is true given the values of its atoms. Logicians can use truth tables to determine whether or not an inference is valid: it is valid if its conclusion must be true given that its premises are true. One of the glories of twentieth century logic was Gödel’s discovery that there are logics in which not all inferences that are valid in their semantic system can be proved using a consistent formal system (see, e.g. Boolos & Jeffrey 1989). (The reason for the stipulation that the system

is consistent is that an inconsistent system would allow any proposition including contradictions to be proved.) The logic of sentential connectives, however, has the happy property that all inferences that are valid on the basis of their truth-functional meanings are also provable in a consistent formal system, and vice versa.

2. The interpretation of connectives in natural language

The psychology of reasoning would be simpler if all connectives in natural language were truth functional. But, temporal connectives, such as “and then” or “before”, are not truth functional. It is true that Bush declared war on terrorism, and that terrorists attacked the USA, but the following assertion is nevertheless false:

Bush declared war on terrorism and then terrorists attacked the USA.

The two events occurred in the opposite order.

In fact, the human interpretative system cannot be truth functional, not even in the case of logical interpretations. As the example in the previous section showed, a truth-functional interpretation starts and ends with truth values. It doesn't take into account what individual atoms mean, what they refer to, or any temporal, spatial, or other such relation between them: all it depends on are truth values. When you understand a sentence, however, you don't end up with its truth value. Indeed, you may never know its truth value, which depends on the relation between what it signifies and the state of the world. Comprehension starts with the construction of the meaning of a sentence; it recovers its referents, their properties, and any relations among them—a process that may depend on knowledge; and it ends with a representation of the possible situations to which the sentence refers. In short, it starts with meanings and ends with models. The moral is clear. No connectives in natural language are interpreted in a truth functional way (see Johnson-Laird & Byrne 2002, Byrne 2005).

Many uses of “if”, “or”, and “and” don't have a logical meaning. The connective and can be interpreted to mean *and then*. The following disjunction:

They played soccer or they played some game
seems innocuous. But, if you learn that the second atom is false, i.e.:

They didn't play any game
you would not infer the truth of the first atom:

They played soccer.

The formal rule I presented earlier would allow this inference to be made, but in real life you wouldn't make it. You know that soccer is a game, and so you interpret the disjunction to be compatible with only two possibili-

ties, and in both of them they played a game (see Johnson-Laird & Byrne 2002, for an account of such “modulations” of interpretation). Hence, the disjunction no longer has a logical meaning.

3. The limitations of working memory

A *mental* model represents a possibility, or, to be precise, the structure and content of the model capture what is common to the different ways in which the possibility could occur—a construal that I owe to a logician, the late Barwise (1993). When you are forced to try to hold in mind several models of possibilities, the task is difficult. To experience this phenomenon of “memory overload” for yourself, try the following problem:

June is in Wales or Charles is in Scotland, or both.

Charles is in Scotland or Kate is in Ireland, or both.

What, if anything, follows?

The disjunctions are inclusive, and so each premise is consistent with three possibilities. The problem, of course, is to combine the two sets of possibilities. In fact, they yield five possibilities, which support the valid conclusion:

June is in Wales and Kate is in Ireland, or Charles is in Scotland, or both.

Five possibilities are too many to hold in mind at the same time, and so, as the theory predicts, this inference is hard. My colleagues and I tested a sample of the general population in an experiment, and only 6% of them drew a valid conclusion (Johnson-Laird et al. 1992). The experiment also examined similar inferences based on exclusive disjunctions:

June is in Wales or Charles is in Scotland, but not both.

Charles is in Scotland or Kate is in Ireland, but not both.

What, if anything, follows?

These premises are compatible with only two possibilities, and they yield the conclusion:

Either June is in Wales and Kate is in Ireland or else Charles is in Scotland.

The problem was easier: 21% of the participants drew this conclusion or an equivalent to it.

Most people go wrong with both sorts of inference, and so you might wonder what conclusions they draw. If they are trying to construct mental models of the various possibilities, then there are two obvious predictions. The first is that if they grasp that there’s more than one possibility but are unable to discern what holds over all of them, then they should respond that there’s no valid conclusion. About a third of responses were of this sort. The second prediction is that if people overlook one or more

possibilities, then their conclusion should correspond to only *some* of the possibilities compatible with the premises. In fact, nearly all of the participants' erroneous conclusions were of this sort. Indeed, the most frequent errors were conclusions based on just a single possibility compatible with the premises. These errors cannot be attributed to blind guessing, because of the improbability of guessing so many conclusions compatible with the premises. People prefer to reason on the basis of a single model. Their erroneous conclusions are so hard to explain if they are relying on formal rules that no-one has so far devised such an explanation (pace Rips 1994, Braine & O'Brien 1998).

A simple way in which to prevent reasoners from being swamped by possibilities is to give them an extra premise that establishes the definite whereabouts of one of the persons, e.g.:

June is in England.

June is in Wales or Charles is in Scotland, but not both.

Charles is in Scotland or Kate is in Ireland, but not both.

What should then happen is that the interpretation of the first two premises yields only a single possibility:

June is in England Charles is in Scotland

The combination of this possibility with those for the third premise yields:

June is in England Charles is in Scotland Kate is *not* in Ireland

In this way, the number of possibilities that have to be kept in mind at any one time is reduced to one. The experiment included some problems of this sort, and they were easy. Diagrams can also improve performance with disjunctive problems, but not just any diagrams. They need to make the task of envisaging alternative possibilities easier (see Bauer & Johnson-Laird 1993).

Your working memory has a limited ability to hold models in mind. A superhuman intelligence, however, wouldn't be limited in this way. Its working memory would not be a bottleneck, and so it could reason with much more complex premises than you can. You don't realize your limitations because your social world is no more complicated than your ability to think about it—it couldn't be—and your reasoning about the physical world is good enough for you to survive.

4. The principle of parsimony

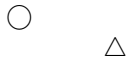
The model theory postulates that mental models are parsimonious. They represent what is possible, but not what is impossible, according to assertions. This principle of parsimony minimizes the load on working memory, and so it applies unless something exceptional occurs to overrule it. It

was introduced in Johnson-Laird & Savary (1999), who referred to it as the principle of “truth”. This name is slightly misleading, and so I have changed it here. Some critics have thought that the principle means that mental models represent only those clauses mentioned in the premises. Such a view, however, would imply wrongly that sentences have the same models regardless of the connectives that occur in them.

The principle of parsimony is subtle because it applies at two levels. At the first level, mental models represent only what is possible. Consider, for example, how they represent the exclusive disjunction:

There is a circle or else there is a triangle but not both.

Its mental models represent the two possibilities:



where “ \bigcirc ” denotes a model of the circle, “ \triangle ” denotes a model of the triangle, and each horizontal line denotes a model of a separate possibility. Hence, the first row in this diagram represents the possibility described in the first clause in the sentence, and the second row represents the possibility described in the second clause. You will notice that two models of possibilities are more parsimonious than the four rows of a truth table, which represent both what is possible and what is impossible according to the premises.

The second level at which the principle of parsimony applies concerns individual models of possibilities: a mental model of a possibility represents a clause in the premises, whether it is affirmative or negative, only when the clause holds in that possibility. This principle is exemplified in the mental models of the disjunction above. The first model represents the possibility of a circle, but not the concurrent impossibility of a triangle. It contains no explicit information about the triangle. Likewise, the second model represents the possibility of a triangle, but not the concurrent impossibility of a circle. It contains no explicit information about the circle.

If you ask people to list what is possible given the preceding exclusive disjunction, they do indeed list a circle as one possibility, and a triangle as another possibility, and they say nothing about the status of the triangle in the first case or the status of the circle in the second case (Johnson-Laird & Savary 1999). Yet, they have not entirely forgotten what is impossible in a possibility that they represent. It is as though they made a mental footnote about it. But, the footnote is soon forgotten if they have to carry out a taxing piece of reasoning or if sentences contain several connectives. Let’s consider a different sentential connective, the conditional, which joins together two clauses using “if” and “then”. Consider the assertion:

If there is a circle then there is a triangle

You might ask: “And if there isn’t circle, what then?” The answer is that there may or may not be a triangle. The conditional in its logical sense

is therefore compatible with three possibilities, which as usual I show on separate lines:

$$\begin{array}{l} \bigcirc \quad \triangle \\ \neg \bigcirc \quad \triangle \\ \neg \bigcirc \quad \neg \triangle \end{array}$$

where “ \neg ” denotes negation. From adolescence or earlier, children list these possibilities, as do adults, when they are asked what is possible given a conditional (see, e.g. Barrouillet & Leças 1999, Barrouillet et al. 2000). However, because it’s difficult to hold them all in mind, when individuals reason from a conditional, they focus on the possibility in which both the “if” clause, the *antecedent*, and the “then” clause, the *consequent*, occur. And so they construct the mental model:

$$\bigcirc \quad \triangle$$

But, if they were to construct only this model, then they would have represented a conjunction: there is a circle *and* there is a triangle. They realize that the antecedent needn’t occur: there needn’t be a circle. But, they defer the construction of an explicit model of this possibility. They construct only a model that has no explicit content. It acts as a “place holder” to remind them that there are other possibilities. The mental models of the conditional are accordingly:

$$\begin{array}{l} \bigcirc \quad \triangle \\ \dots \end{array}$$

where the ellipsis denotes the implicit model. Individuals should make a mental footnote that the possibilities represented in the implicit model are those in which the antecedent doesn’t occur, i.e., there isn’t a circle. If they retain this footnote, then they can flesh out their mental models into *fully explicit* models of the three possibilities. Now, you can understand why there is a difference in difficulty between the two conditional inferences with which I began the chapter. The easy inference follows at once from the mental models of the conditional, whereas the difficult inference does not. One way to make the difficult inference is to flesh out the mental models into fully explicit models; another way, which I will describe presently, is to make a supposition.

Just as there are two sorts of logical disjunction, inclusive and exclusive, so there are two sorts of logical conditional. You may have understood the conditional above to mean that if, *and only if*, there’s a circle then there’s a triangle. This interpretation is known as a biconditional, because it is equivalent to the assertion of two conditionals:

If there is a circle then there is a triangle, and if there isn’t a circle then there isn’t a triangle.

The biconditional is compatible with only two possibilities:

$$\begin{array}{l} \bigcirc \quad \triangle \\ \neg \bigcirc \quad \neg \triangle \end{array}$$

But, it has the same mental models as the regular conditional, except that the footnote states that the implicit model represents the possibility in which both clauses fail to hold. If you retain the footnote, then you should be able to flesh out your mental models into fully explicit models of the two possibilities. One reason that you will try to do so is if you are unable to draw a conclusion from your mental models.

Table 2 summarizes the mental models and the fully explicit models of sentences based on the *logical* meanings of the five principal sentential connectives. The ellipses represent implicit models, which serve as place holders representing other possibilities that as yet have no explicit content and that are constrained by mental footnotes. The fully explicit models flesh out mental models to represent all the clauses in the premises in all the possibilities.

5. Truth tables versus models

You should now understand the difference between truth tables and models. Truth tables represent truth values. Models represent possibilities. For example, the conjunction:

There is *not* a circle and there is a triangle

is represented by a truth table with four rows, which represents whether the atomic propositions are true or false. The only row for the conjunction that is true states in effect:

It is false that there is a circle and it is true that there is a triangle.

In contrast, the conjunction has a single mental model of a possibility:

$\neg \bigcirc \quad \triangle$

Truth values are not possibilities, and the distinction matters in logic. When individuals refer to what is “true” or “false”, or mentally represent these terms, they are at risk of paradox, as in famous example from twentieth century logic:

This sentence is false.

If this sentence is true then it is false; if it is false then it is true. Of course, the sentence seems silly because it has no topic other than itself. Yet, logicians go to any lengths to avoid such paradoxes, because they are a symptom of an inconsistent system (see, e.g. Barwise & Etchemendy 1987). No risk of paradox occurs in referring to possibilities, e.g.:

This sentence is impossible.

The sentence merely makes a false claim about the grammar of English: “true” and “false” refer to the truth values of sentences, but “impossible” does not.

The difference between truth values and possibilities matters in psychology,

Table 2

The mental models and the fully explicit models for five sentential connectives

Connectives	Mental models	Fully Explicit models
<i>Conjunction:</i>		
A and B:	A B	A B
<i>Exclusive disjunction:</i>		
A or B but not both:	A	A \neg B
	B	\neg A B
<i>Inclusive disjunction:</i>		
A or B or both:	A	A \neg B
	B	\neg A B
	A B	A B
<i>Conditional:</i>		
If A then B:	A B	A B
	. . .	\neg A B
		\neg A \neg B
<i>Biconditional:</i>		
If and only if A then B:	A B	A B
	. . .	\neg A \neg B

Key: " \neg " symbolizes negation, and ". . ." a wholly implicit model.

because individuals respond differently to questions about truth and falsity than to questions about possibility and impossibility. For example, they tend to think that conditionals are *true* only in the case that both their clauses are true, but they are happy to list as *possible* all three cases in Table 2, corresponding to fully explicit models. Judgments of truth and falsity call for relating mental models to external possibilities in order to derive truth values. When individuals list possibilities, however, they have only to understand a sentence, and so they can flesh out their mental models into the three fully explicit models of a conditional.

6. Mechanisms of model building

The model theory postulates that humans have a natural disposition to think of possibilities. Alternative possibilities are represented as disjunctions of possibilities; and each model of a possibility represents a conjunction of affirmative and negative propositions. The theory as it applies to logical connectives therefore takes negation, conjunction, and inclusive disjunction, as fundamental. In this second part of the chapter, I am going to describe the mechanisms that construct models. These mechanisms have all been implemented in a computer program, and the program yields a surprising consequence, which I'll get to by and by. But, I begin with negation, and then proceed to connectives.

Here is a problem that turns out to be harder than it seems at first sight (see Barres & Johnson-Laird 2003). List the possibilities given the following assertion: It is not the case both that there is a circle and that there is a triangle. Why isn't the task trivial? The answer is that you don't know the answer, and so you have to infer it. You first have to work out what the unnegated sentence means:

There is a circle and there is a triangle.

It allows just one possibility:

○ △

The negative sentence rules out this possibility to leave its complement, which is all the other possible models based on the same two atoms and their negations. The first one that you're likely to think of is the mirror image of the preceding possibility:

¬○ ¬△

Some individuals go no further, but you will realize that there are two other possibilities, in which one or other of the two shapes is missing:

¬○ △
○ ¬△

In general, the way to infer the correct interpretation of a negative sentence is to take its atoms, and to work out all the possible combinations of them and their negations. You remove from these combinations those that are compatible with the unnegated sentence, and what remains is the answer: the possibilities compatible with the negative sentence. No wonder that people do not cope with the negation of compound sentences well. They tend to be better at negating a disjunction than a conjunction, perhaps because the former yields fewer models than the latter.

Individuals represent a set of alternative possibilities as a list of alternative models. Such a list corresponds to an inclusive disjunction. To combine two such sets of models according to any logical relation between them, calls only for negation, which I've described, and logical conjunction, which I'm

about to describe. When individuals interpret a set of premises, however, they construct a model of an initial clause or premise, and then update this model from the remaining information in the premises.

Let's consider a pair of premises that illustrate the main principles of conjunction:

If there is a triangle then there is a diamond.

There is a circle or else there is a triangle but not both.

Before I tell you what the resulting models are, you might like to think for yourself what possibilities are compatible with the two premises. Most people think that there are two: a triangle and a diamond, or a circle.

The mental models of the first premise are:

\triangle \diamond
 . . .

The core of the interpretative process is to update these models by forming a conjunction of them with the models of the second premise. One possibility according to the second premise is that there is a circle, and so the system conjoins:

\triangle \diamond and \circ

The triangle in the first model here occurs elsewhere in the models containing the circle, and so the interpretative system takes the absence of the triangle from the model containing the circle to mean that there is not a triangle. In effect, the conjunction becomes:

\triangle \diamond and \circ $\neg \triangle$

Because there is now a contradiction—one model contains a triangle and the other its negation—the result is a special null model (akin to the empty set), which represents propositions that are contradictory. It represents what is impossible. The conjunction therefore yields the null model:

nil

The system now conjoins the pair:

\triangle \diamond and \triangle

The diamond doesn't occur elsewhere in the set of models containing the model of the triangle alone, and so the two models are compatible with one another. Their conjunction yields:

\triangle \diamond

Similarly, the conjunction:

. . . and \circ yields \circ

because the circle doesn't occur in the models containing the implicit model. The final conjunction:

. . . and \circ yields nil

because the triangle does occur elsewhere in the models containing the implicit model, and so its absence in the implicit model is treated as akin to its negation. The mental models of the conjunction of the premises are accordingly:

Table 3

The mechanisms for conjoining pairs of mental models and pairs of fully explicit models

1. If one model contains a representation of a proposition, A, which is not represented in the other model, then consider the set of models of which this other model is a member. If A occurs in at least one of these models, then its absence in the current model is treated as its negation (go to mechanism 2); otherwise its absence is treated as its affirmation (go to mechanism 3). This mechanism applies only to mental models.
2. The conjunction of a pair of models containing respectively a proposition and its negation yield the null model, e.g.:
 $A \ B$ and $\neg A \ B$ yield nil.
3. The conjunction of a pair of models that are not contradictory yields a model containing all the elements of both models, e.g.:
 $A \ B$ and $B \ C$ yield $A \ B \ C$.
4. The conjunction of a null model with any model yields the null model, e.g.:
 $A \ B$ and nil yield nil.



I have not shown the null models, because they do not represent possibilities. The two models of possibilities yield the valid conclusion:

There is a triangle and a diamond, or else there is a circle.

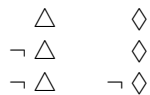
Table 3 summarizes the mechanisms for forming conjunctions of pairs of models.

The same mechanisms apply to the conjunction of fully explicit models. Here are the previous premises again:

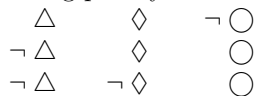
If there is a triangle then there is a diamond.

There is a circle or else there is a triangle but not both.

Their mental models can be fleshed out to be fully explicit by a mechanism that uses mental footnotes, but I'll spare you the details. The fully explicit models of the conditional (see Table 1) are:



Because the disjunction has two models, there are six pair-wise conjunctions, but three of them are contradictions yielding the null model. The remaining pairs yield the following results:



The same conclusion follows as before:

There is a triangle and a diamond or else there is a circle.

But, reasoners who rely mental models will fail to think about the second of these possibilities. They should think that it is impossible to have the diamond and the circle. This prediction is typical of the model theory.

You can make suppositions when you reason, i.e., assumptions for the sake of argument (see, e.g. Byrne et al. 1995). Given a disjunction, such as:

There is a triangle on the board or there is a circle, or both.

you can make the supposition that there isn't a triangle on the board, and then infer as a consequence that in that case there is a circle on the board. You hold in mind a possibility, which in this case corresponds to the negation of an atom in the premise, and then treat it as though it was asserted categorically. You can then use the inferential mechanisms that I have already described. If you are prudent, you remember that any conclusion depends on a supposition, and take this fact into account in formulating a final conclusion. If a supposition leads to a contradiction (the null model), some individuals appreciate that the supposition is impossible granted the truth of the premises. The procedure is identical to the one that occurs in the construction of models of the following sort of conditional:

If A then both B and not B.

The conjunction, *B and not B*, yields the null model. The interpretation of the conditional calls for the conjunction of A and *nil*, which yields *nil* (see Table 3). What happens then depends on whether individuals are relying on mental models or fully explicit models. With mental models, there remains only the implicit model, which yields no conclusion. But, the fully explicit models of the conditional are:

A	nil
¬ A	nil
¬ A	¬ nil

The negation of *nil* in the third model yields the disjunction of the atoms that led to its construction, and so this conjunction yields the conclusion:
not A.

The corresponding principle in logic is known as *reductio ad absurdum*. In the model theory, it is a consequence of a mechanism that makes suppositions, and of reasoning from fully explicit models.

In a review of theories of conditionals, Evans & Over (2004) claimed that the model theory makes no use of suppositions, despite our several papers to the contrary (e.g. Byrne et al. 1995). They also argued that the model theory is truth functional, despite the arguments that I have summarized above. Their review is otherwise valuable. It is a pity that they mangle the model theory so badly, because it makes sense of phenomena that are otherwise puzzling for them, e.g., the difference that I described earlier between judgments of truth value and the listing of possibilities.

7. Superhuman reasoning

A computer program that I wrote to simulate the model theory can make inferences that are far beyond the ability of human reasoners working without benefit of logic. Only a superhuman intelligence, such as Hercule Poirot (Agatha Christie's famous fictional detective), could solve the following problem without paper and pencil:

Who helped to murder Mr. Ratchett on the Orient Express?
 If Pierre helped if Dr. Constantine did then Greta helped too.
 If not both Harriet and Hector helped then Dr. Constantine didn't help.
 Greta didn't help or Dr. Constantine did.
 Harriet didn't help or the Princess Drago-Miroff, Mary, and Colonel Arbuthnot all helped.
 If Hector or Mary helped then Pierre helped or Colonel Arbuthnot didn't help.
 So, who helped, who didn't help, and for whom is it impossible to say?

There are eight atomic propositions in the premises, and so their truth table has 256 rows. Likewise, there are multiple models, but if you build them up premise by premise, the final result is a single model. It shows that all eight individuals helped to commit the murder. In many other inferences, of course, the premises yield multiple models, but an algorithm exists for drawing parsimonious conclusions that describe them (see Johnson-Laird & Byrne 1991, Ch. 9).

8. Some illustrative inferences

To illustrate the model theory and its predictions, I am going to consider some inferences that human reasoners *can* make. The first inference is:

Either Jane is kneeling by the fire and she is looking at the TV or else Mark is standing at the window and he is peering into the garden.

Jane is kneeling by the fire.

Does it follow that she is looking at the TV?

Most people say: "yes" (Walsh & Johnson-Laird 2004). A second inference has the same initial premise, but it is followed instead by the categorical denial:

Jane is not kneeling by the fire.

and the question is:

Does it follow that Mark is standing at the window?

Again, most individuals say: “yes”. Let’s see what the theory predicts. The first premise in both inferences is the same exclusive disjunction of two conjunctions. The theory predicts that individuals should rely on mental models. Hence, they should interpret the first conjunction, Jane is kneeling by the fire and she is looking at the TV, and build a model representing this possibility, which I will abbreviate as follows:

Jane: kneeling looking

They should build an analogous model of the second conjunction:

Mark: standing peering

These two models must now be combined according to an exclusive disjunction. An exclusive disjunction has two mental models, which represent the two conjunctions only in the possibilities in which they hold:

Jane: kneeling looking

Mark: standing peering

For the first inference, the conjunction of the categorical premise:

Jane is kneeling

with the first model of the disjunction yields:

Jane: kneeling looking

Its conjunction with the second model of the disjunction yields the null model. Hence, the premises yield only the model:

Jane: kneeling looking

and so individuals should respond: yes, Jane is looking at the TV. This analysis may strike you as obvious.

In fact, the inference is a fallacy. The principle of parsimony postulates that individuals normally represent what is possible, but not what is impossible. When I first wrote the computer program to simulate the theory, and inspected its output for a certain problem, I thought that there was a bug in the program. I searched for the bug for half a day, before I realized that the program was correct, and the error was in my thinking. What the program revealed is the discrepancy between mental models and fully explicit models. The theory therefore predicted that individuals should reason in a fallacious way for certain inferences. Indeed, the fallacies turn out to be so compelling in some cases that they resemble cognitive illusions, and so my colleagues and I refer to them as “illusory” inferences.

If you succumbed to the illusion, then you are in the company of Clare Walsh and myself. We studied these inferences, but it took us a couple of days to realize that they were illusory, and that was *after* the discovery of other sorts of illusions. The fully explicit models of the exclusive disjunction reveal the correct conclusion:

Jane: kneeling looking Mark: \neg standing \neg peering

Jane: kneeling looking Mark: \neg standing peering

Jane: kneeling looking Mark: standing \neg peering

Jane: \neg kneeling \neg looking Mark: standing peering

Jane:	\neg kneeling	looking	Mark:	standing	peering
Jane:	kneeling	\neg looking	Mark:	standing	peering

When one conjunction is true, the other conjunction is false, and you will remember from my earlier account that there are three ways in which a conjunction can be false. The categorical premise that Jane is kneeling rules out the fourth and fifth possibilities. But, contrary to the illusory inference, it leaves one possibility—the sixth one—in which Jane is kneeling but not looking at the TV. That is why the illusory inference is invalid. Granted that Jane is kneeling, it does not follow that she is looking at the TV.

The second problem that I described has the categorical premise that Jane is not kneeling by the fire, and poses the question of whether it follows that Mark is standing by the window. Most people respond, “yes”, which is a conclusion supported by the mental models shown above. The fully explicit models show that this inference *is* valid. The categorical premise eliminates all but the fourth and fifth models, and in both of them Mark is standing by the window. Our main experiment examined a series of illusory inferences and control problems of this sort. The participants were much more likely to respond correctly to the control problems (78% correct) than to the illusory problems (10% correct): 34 of the 35 participants showed this difference.

Illusory inferences occur in many domains, including reasoning with quantifiers (Yang & Johnson-Laird 2000*a,b*), deontic reasoning (Bucciarelli & Johnson-Laird 2005), and assessing whether or not sets of assertions are consistent (Johnson-Laird et al. 2004). I will describe two more examples. The first example (from Goldvarg & Johnson-Laird 2000) calls for reasoning about what is possible:

Only one of the following premises is true about a particular hand of cards:

- There is a king in the hand or there is an ace, or both.
- There is a queen in the hand or there is an ace, or both.
- There is a jack in the hand or there is a 10, or both.

Is it possible that there is an ace in the hand?

The model theory postulates that individuals consider the possibilities for each of the three premises. That is, they assume that the first premise is the one that is true, and consider the consequences; then they assume that the second premise is the one that is true and consider the consequences, and then they assume that the third premise is the one that is true and consider the consequences. However, because the question asks only whether an ace is *possible*, they can stop as soon as they find a premise that allows the presence of the ace in the hand. What is wrong with this procedure? The answer is that when individuals consider the truth of one premise, they should also consider the concurrent *falsity* of the other two premises. But, that is exactly what the principle of parsimony predicts they will not do.

For the first premise, they accordingly consider three models, which each correspond to a possibility given the truth of the premise:

king	ace
king	ace

Two of the models show that an ace is possible. Hence, on the basis of this premise alone individuals should respond, “yes”. The second premise supports the same conclusion. The third premise is compatible with it. In fact, it is an *illusion of possibility*: reasoners infer wrongly that a card is possible. If there were an ace, then two of the premises would be true, contrary to the rubric that only one of them is true. The same strategy, however, yields a correct response to a control problem in which only one premise refers to an ace. A problem to which reasoners should respond “no”, and thereby succumb to an *illusion of impossibility*, can be created by replacing the two occurrences of “there is an ace” in the premises above with, “there is not an ace”. Its control problem contains only one premise with the clause, “there is not an ace”.

Figure 1 presents the results of an experiment in which we gave students 16 inferences, four of each of the four sorts. Half of the illusions were based on disjunctive premises, and half were based on conditionals. The participants’ confidence in their conclusions did not differ reliably from one sort of problem to another. As the Figure shows, they were very susceptible to the illusions but performed well with the control problems, and the illusions of possibility were more telling than those of impossibility. To infer that a situation is impossible calls for a check of every model, whereas to infer that a situation is possible does not, and so reasoners are less likely to make the inference of impossibility. This difference also occurs in problems that are not illusory (Bell & Johnson-Laird 1998).

With hindsight, it is surprising that nearly everyone responded “yes” to the first of the problems above, because it seems obvious that an ace renders two of the premises true. We therefore carried out a replication with two groups of participants, and half way through the experiment, we told one group to check whether their conclusions met the constraint that only one of the premises was true. This procedure had the advantage that the participants did not have to envisage the circumstances in which the premises did not hold. The group that received the special instruction was thereafter much less likely to commit the fallacies (Goldvarg & Johnson-Laird 2000). If the preceding illusions result from a failure to reason about what is false, then any manipulation that emphasizes falsity should reduce them. The rubric, “Only one of the following two premises is *false*,” did reduce their occurrence (Tabossi et al. 1998), as did the prior production of false instances of the premises (Newsome & Johnson-Laird 1996).

The second example of an illusion is very compelling. The rubric, “one of

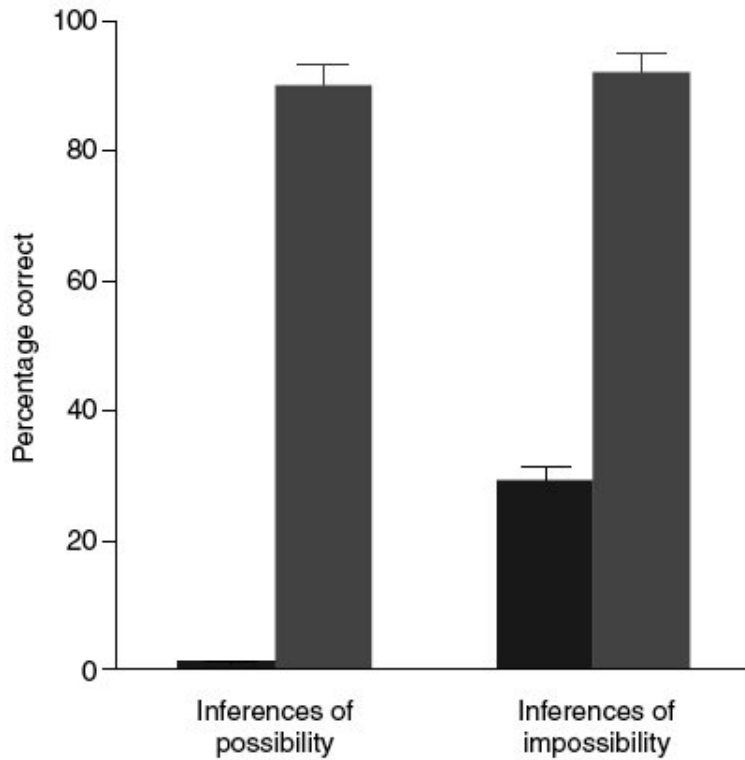


Fig. 1. The percentages of correct responses to fallacious inferences that are illusory and their control problems (based on Goldvarg and Johnson-Laird, 2000).

these assertions is true and one of them is false”, is equivalent to an exclusive disjunction between two assertions. Consider this problem, which is based on an exclusive disjunction:

Suppose you know the following about a particular hand of cards:

If there is a jack in the hand then there is a king in the hand, or else if there isn't a jack in the hand then there is a king in the hand.

There is a jack in the hand.

What, if anything, follows?

Nearly everyone—experts and novices alike—infer that there is a king in the hand (Johnson-Laird & Savary 1999). Yet, it is a fallacy granted a disjunction, exclusive or inclusive, between the two conditional assertions. The disjunction entails that one or other of the two conditionals could be false; and if one of them is false, then there may not be a king in the hand.

Suppose, for instance, that the first conditional is false. There could then be a jack but *not* a king—a judgment with which most individuals concur (see, e.g. Oaksford & Stenning 1992). And so the inference that there is a king is invalid: the conclusion could be false.

An experiment examined the preceding problem and another illusion, and compared them with two control problems in which the neglect of false cases should not impair performance (Johnson-Laird & Savary 1999). The participants committed both fallacies in 100 percent of cases, and yet drew valid inferences for the control problems in 94 percent of cases. The participants were again confident in both their illusory conclusions and their correct control conclusions.

Because so many expert psychologists have succumbed to illusory inferences, we have accumulated many putative explanations for them. For example, the premises may be so complex, ambiguous, or odd, that they confuse people, who, as a result, commit a fallacy. This hypothesis overlooks the fact that the participants are very confident in their conclusions, and that the control problems are equally complex. Likewise, when the illusions and controls are based on the *same* premises, but different questions in the form of conjunctions, participants still commit the fallacies and get the control problems correct (Goldvarg & Johnson-Laird 2000).

Other putative explanations concern the interpretation of conditionals. Individuals make the illusory inference with problems of this sort:

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

If there is a jack then there is a king.

If there isn't a jack then there is a king.

This assertion is definitely true:

There is a jack.

Naïve individuals understand that a conditional, such as:

If there is jack then there is a king

is false in the case that there is jack but not a king. They also understand that the rubric to this problems mean that one conditional is true and the other conditional is false. Hence, on their own account they should refrain from inferring that there is a king. The analysis depends on nothing else. However, even if some special factors exacerbate the illusions with conditionals, other illusions occur with problems that do not contain conditionals, such as the problem with which I started this section of the chapter.

Many other robust phenomena in reasoning appear to arise from the principle of parsimony and the resulting neglect of what is impossible or false. They include the results of Wason's "selection" task in which individuals fail to grasp the relevance of an instance of a false consequent to testing the truth or falsity of a conditional (see, e.g. Wason 1966, Wason & Johnson-Laird 1972).

9. Conclusions

This chapter has explained the mechanisms that construct models based on the logical interpretation of connectives. These models do not represent truth values, but sets of possibilities. Individuals adopt a variety of strategies to cope with reasoning problems, e.g., they may be guided by a given conclusion, they may work forwards from the premises, they may make a supposition, and so on (Van der Henst et al. 2002, Johnson-Laird & Hasson 2003). But, regardless of strategy, inferences that depend on a single model are easier than those that depend on multiple models.

Mental models abide by the principle of parsimony: they represent only possibilities compatible with the premises, and they represent clauses in the premises only when they hold in a possibility. Fully explicit models represent clauses when they do not hold too. The advantage of mental models over fully explicit models is that they contain less information, and so they are easier to work with. But they can lead reasoners astray. The occurrence of these systematic and compelling fallacies is shocking. The model theory predicts them, and they are a “litmus” test for mental models, because no other current theory predicts them. They have so far resisted explanation by theories of reasoning based on formal rules of inference, because these theories rely on valid rules. For several years, my former colleague Yingrui Yang has sought an explanation based on a revised formal rule theory, but he has yet to succeed. To reason only about what is possible is a sensible way to cope with limited processing capacity, but it does lead to illusions. Yet, it does not imply that people are irredeemably irrational. The fallacies can be alleviated with preventative methods. Otherwise, however, reasoners remain open to the illusion that they grasp what is in fact beyond them.

References

- Barres, P. & Johnson-Laird, P. (2003), ‘On imagining what is true (and what is false)’, *Thinking & Reasoning* **9**, 1–42.
- Barrouillet, P., Grosset, N. & Leças, J. F. (2000), ‘Conditional reasoning by mental models: chronometric and developmental evidence’, *Cognition* **75**, 237–266.
- Barrouillet, P. & Leças, J.-F. (1999), ‘Mental models in conditional reasoning and working memory’, *Thinking & Reasoning* **5**, 289–302.
- Barwise, J. (1993), ‘Everyday reasoning and logical inference’, *Behavioral and Brain Sciences* **16**, 337–338.
- Barwise, J. & Etchemendy, J. (1987), *The Liar: An Essay in Truth and Circularity*, Oxford University Press, New York.

- Bauer, M. & Johnson-Laird, P. (1993), 'How diagrams can improve reasoning', *Psychological Science* **4**, 372–378.
- Bell, V. & Johnson-Laird, P. (1998), 'A model theory of modal reasoning', *Cognitive Science* **22**, 25–51.
- Berry, G. (1952), Peirce's contributions to the logic of statements and quantifiers, in P. Wiener & F. Young, eds, 'Studies in the Philosophy of Charles S. Peirce', Harvard University Press, Cambridge, MA.
- Boolos, G. & Jeffrey, R. (1989), *Computability and Logic*, 3 edn, Cambridge University Press, Cambridge.
- Braine, M. & O'Brien, D., eds (1998), *Mental Logic*, Erlbaum, Mahwah, NJ.
- Bucciarelli, M. & Johnson-Laird, P. (2005), 'Naïve deontics: a theory of meaning, representation, and reasoning', *Cognitive Psychology* (in press).
- Byrne, R. (2005), *The Rational Imagination: How People Create Alternative to Reality*, MIT Press, Cambridge, MA.
- Byrne, R., Handley, S. & Johnson-Laird, P. (1995), 'Reasoning from suppositions', *Quarterly Journal of Experimental Psychology* **48A**, 915–944.
- Evans, J., Newstead, S. & Byrne, R. (1993), *Human Reasoning: The Psychology of Deduction*, Erlbaum, Hillsdale, NJ.
- Evans, J. & Over, D. (2004), *If*, Oxford University Press, Oxford.
- Garnham, A. (2001), *Mental Models and the Representation of Anaphora*, Psychology Press, Hove, East Sussex.
- Gentner, D. & Stevens, A., eds (1983), *Mental Models*, Erlbaum, Hillsdale, NJ.
- Goldvarg, Y. & Johnson-Laird, P. (2000), 'Illusions in modal reasoning', *Memory & Cognition* **28**, 282–294.
- Jeffrey, R. (1981), *Formal Logic: Its Scope and Limits*, 2 edn, McGraw-Hill, New York.
- Johnson-Laird, P. (1983), *Mental Models: Towards a Cognitive Science of Language, Inference, and Consciousness*, Cambridge University Press/Harvard University Press, Cambridge/Cambridge, MA.
- Johnson-Laird, P. & Byrne, R. (1991), *Deduction*, Erlbaum, Hillsdale, NJ.
- Johnson-Laird, P. & Byrne, R. (2002), 'Conditionals: A theory of meaning, pragmatics, and inference', *Psychological Review* **109**, 646–678.
- Johnson-Laird, P., Byrne, R. & Schaeken, W. (1992), 'Propositional reasoning by model', *Psychological Review* **99**, 418–439.
- Johnson-Laird, P., Girotto, V. & Legrenzi, P. (2004), 'Reasoning from inconsistency to consistency', **111**, 640–661.
- Johnson-Laird, P. & Hasson, U. (2003), 'Counterexamples in sentential reasoning', *Memory & Cognition* **31**, 1105–1113.
- Johnson-Laird, P., Legrenzi, P., Girotto, V., Legrenzi, M. & Caverni, J.-P. (1999), 'Naïve probability: a mental model theory of extensional reasoning', *Psychological Review* **106**, 62–88.
- Johnson-Laird, P. & Savary, F. (1999), 'Illusory inferences: A novel class of erroneous deductions', *Cognition* **71**, 191–229.
- Marr, D. (1982), *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*, W.H. Freeman, San Francisco.
- Medvedev, Z. A. (1990), *The Legacy of Chernobyl*, W.W. Norton, New York.

- Metzler, J. & Shepard, R. (1982), Transformational studies of the internal representations of three-dimensional objects, *in* R. Shepard & L. Cooper, eds, 'Mental Images and Their Transformations', MIT Press, Cambridge, MA, pp. 25–71.
- Newsome, M. & Johnson-Laird, P. (1996), An antidote to illusory inferences, *in* 'Proceedings of the Eighteenth Annual Conference of the Cognitive Science Society', Erlbaum, Mahwah, NJ, p. 820.
- Oakhill, J. & Garnham, A., eds (1996), *Mental Models in Cognitive Science*, Psychology Press, Hove, Sussex.
- Oaksford, M. & Stenning, K. (1992), 'Reasoning with conditionals containing negated constituents', *Journal of Experimental Psychology: Learning, Memory, and Cognition* **18**, 835–854.
- Polk, T. & Newell, A. (1995), 'Deduction as verbal reasoning', *Psychological Review* **102**, 533–566.
- Rips, L. (1994), *The Psychology of Proof*, MIT Press, Cambridge, MA.
- Stevenson, R. (1993), *Language, Thought and Representation*, Wiley, New York.
- Tabossi, P., Bell, V. & Johnson-Laird, P. (1998), Mental models in deductive, modal, and probabilistic reasoning, *in* C. Habel & G. Rickheit, eds, 'Mental Models in Discourse Processing and Reasoning', John Benjamins, Berlin.
- Van der Henst, J.-B., Yang, Y. & Johnson-Laird, P. (2002), 'Strategies in sentential reasoning', *Cognitive Science* **26**, 425–468.
- Walsh, C. & Johnson-Laird, P. (2004), 'Co-reference and reasoning', *Memory & Cognition* **32**, 96–106.
- Wason, P. (1966), Reasoning, *in* B. Foss, ed., 'New Horizons in Psychology', Penguin, Harmondsworth, Middx.
- Wason, P. & Johnson-Laird, P. (1972), *The Psychology of Deduction: Structure and Content*, Harvard University Press/Batsford, Cambridge, MA/London.
- Wittgenstein, L. (1922), *Tractatus Logico-Philosophicus*, Routledge & Kegan Paul, London.
- Yang, Y. & Johnson-Laird, P. (2000a), 'How to eliminate illusions in quantified reasoning', *Memory & Cognition* **28**, 1050–1059.
- Yang, Y. & Johnson-Laird, P. (2000b), 'Illusory inferences with quantified assertions: How to make the impossible seem possible, and vice versa', *Memory & Cognition* **28**, 452–465.