Journal of Experimental & Theoretical Artificial Intelligence
Vol. 18, No. 2, June 2006, 121–148

Models and heterogeneous reasoning

P. N. JOHNSON-LAIRD*

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

(*Received 9 June 2003; in final form 24 June 2005*)

Human reasoning is heterogeneous; it is based on information from perception, discourse, and knowledge. This paper outlines a theory that shows how these diverse sources of information are integrated, and how they can yield necessary, possible, and probable conclusions. At the heart of the theory is the notion of a mental model. The paper shows how the theory works for spatial reasoning. It extends the theory to sentential reasoning, and it corroborates the theory's predictions about the use of diagrams to facilitate reasoning. Finally, it draws some conclusions about heterogeneous reasoning.

Keywords: Human reasoning; Heterogeneous reasoning; Mental models; Spatial reasoning; Sentential reasoning

1. Introduction

Reasoning in daily life depends on information from various sources. You look out of the window and see that is raining. You know that you have to go out. You also know that if you go out without an umbrella then you will get wet. You ask your spouse for the umbrella, and your spouse tells you: 'It's lost'. And so, you infer, you'll get wet. Therefore the mental processes for reasoning must engage with heterogeneous premises. One feasible method is to translate them all into expressions in a mental language to which inferential processes apply. This procedure is necessary if reasoning depends on formal rules of inference akin to those of logic, a view that has long had its proponents in cognitive science (Inhelder and Piaget 1958, Osherson 1974–1976, Smith *et al*. 1992, Rips 1994, Braine and O'Brien 1998, Nisbett 2003). Yet, there are grounds for supposing that this view may be mistaken, and the present paper describes an alternative theory of how naive individuals reason from heterogeneous premises. By 'naive' we mean those who have received no

*Email: phil@princeton.edu

Journal of Experimental and Theoretical Artificial Intelligence
ISSN 0952–813X print/ISSN 1362–3079 online © 2006 Taylor & Francis
http://www.tandf.co.uk/journals
DOI: 10.1080/09528130600558091

training in logic. It is a striking fact, at least to some logicians, that such individuals are able to make valid deductions, but how they do so is a matter of long-standing controversy in psychology.

The plan of the paper is straightforward. It begins with the alternative theory based on mental models. It illustrates how the theory works for the domain of spatial reasoning. It shows how the theory extends to other sorts of reasoning. It describes how diagrams, as the theory predicts, can facilitate reasoning. Finally, it draws some conclusions about heterogeneous reasoning.

2. The theory of mental models

The concept of mental models has a long history. In a seminal book, the Scottish psychologist and physiologist Kenneth Craik (1943) wrote:

> If the organism carries a 'small-scale model' of external reality and of its own possible actions within its head, it is able to try out various alternatives, conclude which is the best of them, react to future situations before they arise, utilize the knowledge of past events in dealing with the present and the future, and in every way to react in a much fuller, safer, and more competent manner to the emergencies which face it.

Craik's view of models has several nineteenth-century antecedents in the work of Kelvin, Boltzmann, and Maxwell (Johnson-Laird 2004). However, the intellectual grandfather of the theory of mental models is the American logician and philosopher Charles Sanders Peirce.

Peirce and Frege independently invented modern logic, i.e. the predicate calculus. This calculus concerns inferences in a formal language containing idealized versions of negation, of sentential connectives such as 'if' and 'or', and of quantifiers such as 'all' and 'some'. Peirce devised two diagrammatic systems for reasoning in the predicate calculus (and in more expressive domains), not to improve reasoning, but to display its underlying steps. He wrote:

> Deduction is that mode of reasoning which examines the state of things asserted in the premisses, forms a diagram of that state of things, perceives in the parts of the diagram relations not explicitly mentioned in the premisses, satisfies itself by mental experiments upon the diagram that these relations would always subsist, or at least would do so in a certain proportion of cases, and concludes their necessary, or probable, truth. (Peirce 1931–1968, Vol. 1, para. 66)

Peirce stressed that diagrams can be **iconic**, i.e. they can have the same structure as what they represent (Peirce 1931–1968, Vol. 4, para. 447). In abstract algebra, this notion of having the same structure is known as a homomorphism (Barwise and Etchemendy 1995). It is the inspection of an iconic diagram that reveals truths other than those of the premises used in its construction (Peirce 1931–1968, Vol. 4, para. 530). On the use of iconic representations to reason, Peirce anticipates in a striking way the modern theory of mental models (Johnson-Laird 1983, pp. 125, 136). So what is this theory?

In a nutshell, it makes five assumptions.

(1) When individuals reason, they envisage what is possible given the premises and their knowledge.

(2) Each possibility that they envisage is represented in a separate mental model, which as far as possible has an iconic structure. Accordingly, the structure of a model corresponds to the structure of what it represents.

(3) Mental models follow a principle of **truth**: a model represents propositions in the premises only if they are true in the possibility that the model represents. The assumption yields a surprising prediction: the neglect of what is false should usually be harmless, but certain inferences should be like illusions, i.e. they will have conclusions that are compelling, yet that are completely wrong. As we shall see, these illusory inferences do occur.

(4) To draw a conclusion, reasoners find a proposition that is not asserted in the premises, but that holds in the models of the premises. As Peirce suggested, they check that it holds in all, some, or a proportion of the models, and they formulate a corresponding conclusion about its necessity, possibility, or probability. The theory accordingly integrates logical and probabilistic reasoning depending on possibilities (Johnson-Laird *et al.* 1999).

(5) Naive reasoners are able to refute an invalid conclusion using a counter-example, i.e. a model of the premises in which the conclusion is false (Johnson-Laird and Hasson 2003). As Quine (1974, p. 75) pointed out, reasoning based on formal rules of inference provides no general way to reach a verdict of invalidity, whereas the construction of a counter-example establishes it at once (see also Barwise 1993). Halpern and Vardi (1991) have championed the same idea for the checking of proofs in artificial intelligence.

The theory has received support from psychological experiments (reviewed by Johnson-Laird 2001). It has also been implemented in computer programs that take verbal premises as their input and evaluate or formulate conclusions. The next section illustrates how the theory works in the domain of spatial reasoning. But, how does visual perception provide premises? The answer to this question derives from the work of the late David Marr and his colleagues.

Marr (1982) argued that vision is an unconscious inference from the images on the retina to a mental model that makes explicit what things are where in the scene. The inference makes use of a series of mental representations. Pure vision, of the sort that evolved in mammals, is initiated in the physical interaction between light focused on the retina and the visual pigment in retinal cells, and focusing ensures that each retinal cell receives light from just one point in the scene. Retinal cells convert light into nerve impulses. Pure vision yields the 'two-and-a-half-dimensional' sketch, which makes explicit the relative distance and orientation (with respect to the observer) of each visible surface in the scene. If you are to move about safely, however, you need to identify the entities in the world and their spatial interrelations. Such a representation is independent of your point of view. When you walk into a building and recognize that it contains walls, doors, and stairs, you can easily find your way to your goal, say a particular room. You can do so, Marr argued, because vision solves three problems: it constructs a mental model that makes explicit the three-dimensional shapes of everything in the scene, it uses these shapes to identify the objects, and it makes explicit their locations in relation to one another.

How the visual system identifies objects is controversial and not well understood. No system for object recognition in artificial intelligence performs on a par with human vision. Marr postulated that the process depends on two steps. First, the visual system represents the shapes of objects in terms of their own canonical axes (e.g. a pencil is a long thin cylinder). Secondly, the system compares such a shape with its mental catalogue of the shapes of all known objects. Each entry in the catalogue is itself a model which decomposes the object into the shapes of its component parts and their interrelations. At the highest level, the gross shape of the object is made explicit, but at lower levels the detailed shapes of its parts are fleshed out. The matching of a percept to a catalogued model is complicated and perhaps computationally intractable. A cue about the shape of an object may access a model in the catalogue, which is then used 'top down' to try to match the rest of the percept. What seems clear, however, is that human vision does yield models of the world, and they can play a part in reasoning along with models that have different origins.

3. Models and spatial reasoning

Humans combine information about spatial relations from perception, description, and knowledge. The thesis of this paper is that this information is integrated in mental models. But, how do humans reason about spatial relations? The present section aims to answer this question, and it will do so by focusing on reasoning from verbal premises, because they have been the focus of the relevant psychological studies. Indeed, most studies have examined very simple problems, such as the following.

The triangle is on the right of the circle.
The square is on the left of the circle.
Therefore, the triangle is on the right of the square.

The orthodox way to make such inferences is to use formal rules of inference and meaning postulates that capture the logical properties of relations. The reasoning system translates the premises into an internal representation of their logical form:

(1) (Right-of triangle circle).
(2) (Left-of square circle).

It elicits a meaning postulate expressing the transitivity of 'Right of':

(3) For any x, any y, and any z, if (Right-of x y) and (Right-of y z) then (Right-of x z) and a meaning postulate relating 'left-of' to its converse 'right-of':

and a meaning postulate relating 'Left-of' to its converse 'Right-of':

(4) For any x, and any y, if (Left-of x y) then (Right-of y x).

The system then uses its formal rules of inference to derive the conclusion from these premises. The first step is to instantiate the variables in the two meaning postulates with the appropriate individuals. Three applications of the rule of 'universal instantiation' to the meaning postulate for transitivity yield:

(5) If (Right-of triangle circle) & (Right-of circle square) then (Right-of triangle square).

Two similar instantiations applied to the meaning postulate interrelating the two relations yield:

(6) If (Left-of square circle) then (Right-of circle square).

The derivation then proceeds using the formal rule of *modus ponens* (*If A then B; A; Therefore, B*):

(7) (Right-of circle square) – from lines 2 and 6.

This interim conclusion can be conjoined with premise 1, using a rule for conjunction (*A; B; Therefore, A & B*):

(8) (Right-of triangle circle) & (Right-of circle square).

Finally, another application of *modus ponens* (to lines 5 and 8) produces:

(9) (Right-of triangle square) – from lines 5 and 8.

This line is the desired conclusion: the triangle is on the right of the square. It seems implausible that so simple an inference, which normally takes people less than 2 seconds to make, is derived in such a long-winded way.

Consider a more complicated inference:

> The black ball is directly beyond the cue ball.
> The green ball is on the right of the cue ball, and there is a red ball between them.
> Therefore, if I move so that the red ball is between me and the black ball, then the cue ball is on my left.

To make such an inference using formal rules of inference would call for meaning postulates to capture the effects of movement. However, human reasoners are more likely to make the inference by constructing a mental model of the scene described in the premises. This model may yield a vivid visual image, but such an experience is irrelevant to inference—a point to which we return later. What is crucial is that the model is iconic, and so the conclusion can be read off from it. The model of the premises depicts the arrangement from the speaker's point of view. This point of view shifts in the conclusion so that the red and black balls are lined up, and the model then yields the conclusion.

In order to explore mental models, the author has implemented a computer program that makes simple spatial inferences from verbal descriptions. The program uses representations of the meanings of spatial relations, and transitivity is an emergent property of these meanings rather than represented in a meaning postulate. The meanings of spatial relations are based on increments to the Cartesian coordinates of three-dimensional models:

	Left–right	Front–back	Up–down
On the right of	1	0	0
On the left of	−1	0	0
In front of	0	1	0
In back of	0	−1	0
Above	0	0	1
Below	0	0	−1

Hence, given a reference entity *B* in a spatial model, the position of another entity *A*, which satisfies the description *A is on the right of B*, is obtained by incrementing the value of the position of *B* on the left–right axis and holding constant the values of the other axes on which *B* is located, i.e. incrementing them by zero. Hence, the meaning of *on the right of* is defined by the triple 1 0 0.

The program has procedures for constructing models, formulating conclusions that hold in models, and testing whether conclusions hold in models. The way that it works can be illustrated using the earlier inference:

The triangle is on the right of the circle.
The square is on the left of the circle.
Therefore, the triangle is on the right of the square.

The program has a parser and a compositional semantics, i.e. a set of semantic procedures associated with each rule in its grammar and each word in its lexicon, which it uses to construct a representation of the *meaning* of each sentence. This so-called 'propositional' representation for the first premise is

((1 0 0) Δ o)

As we saw, the parameters (1 0 0) specify which axes of the spatial model need to be incremented in order to put the triangle, represented as Δ, into the model so that it is on the right of the circle, represented as o. The value 1 in (1 0 0) calls for the value of the circle on the right–left axis to be incremented, i.e. the program keeps adding 1 to this value until it finds a position in the model that the triangle can occupy. The zeros in (1 0 0) call for the values of the circle on the other two axes (front–back and up–down) to be held constant in putting the triangle into the model. No prior models of the discourse exist, because this premise is the first. Hence a procedure is called that uses this propositional representation to build a minimal spatial model:

o Δ

where the horizontal dimension corresponds to the left-to-right dimension in the situation. Such a diagram depicts a mental model, and these diagrams will often be referred to as though they were mental models. Each token in the mental model has a property corresponding to the shape of the entity it represents, and the two tokens are in a spatial relation corresponding to the relation between the two entities in the situation described by the assertion. Therefore the model is iconic. A crucial feature of such a spatial model is that elements in the model can be accessed and updated in terms of parameters corresponding to axes.

The interpretation of the second premise:

The square is on the left of the circle.

yields the propositional representation

((− 1 0 0) □ o)

This representation contains an referent, o, that is already represented in the initial model, and so a procedure is called that uses the representation to update this model by adding the square in the appropriate position:

□ o Δ

Reasoning based on models is different from reasoning based on formal rules of inference. A conclusion is necessary given the truth of the premises if the conclusion must be true too. Therefore what is needed is a model-based method to test for this condition. Assertions can be true in indefinitely many different situations, and so it is out of the question to test that a conclusion holds true in all of them. However, it can be done in certain domains precisely because a mental model, as Barwise (1993) pointed out, can stand for indefinitely many possibilities—all those that have in common what the model represents. For the putative conclusion in the example

The triangle is on the right of the square.

Both referents in its propositional representation occur in an existing model. Entities can sometimes be in different existing models, and then a procedure is called to integrate the models according to the premise. However, in the present case the two entities are in the same model, and so a procedure is called to verify whether the propositional representation holds in this model. This procedure returns the value *true*, and this value elicits a further procedure that searches for an alternative model of the premises in which the conclusion is false. The search fails, and so the conclusion is evaluated as valid. If the verification procedure returns the value *false*, then this value elicits a further procedure that searches for an alternative model of the premises in which the conclusion is true. If it succeeds, then the premise is consistent with what went before; if it fails, then the premise is inconsistent with what went before. In this way, the program makes non-monotonic inferences, and it can handle premises that are referentially indeterminate. It constructs the most convenient model (as it scans a model to find a place in which to insert an item), and if the model is wrong, then it can be modified to take into account a premise that eliminates the indeterminacy.

Why not construct models directly from statements rather than through the intermediary of propositional representations, which capture their meanings? Why not, in terms of philosophical jargon, construct the *extensional* representation directly without the *intensional* representation? In fact, when a procedure searches for an alternative model of the premises to refute a conclusion, it needs access to the meaning of the premises other than the model itself. Consider, for example, the following model of a spatial arrangement

□ o Δ

and the putative conclusion

The square is on the left of the circle.

The conclusion is true in the model, but the system cannot determine whether it follows from the premises unless it has access to some other representation of the premises. Suppose that the premises were, in fact,

The circle is on the left of the triangle.
The triangle is on the right of the square.

In this case, the following model is also consistent with the premises, but refutes the conclusion:

o □ Δ

In other words, a model alone does not allow the premises to be reconstructed in a unique way. Hence deduction calls for an independent access to the premises, and a representation of their meanings provides this essential information.

The program has no need for meaning postulates capturing transitivity or other logical properties. As the preceding inference shows, they emerge from the meaning of the relation and how it is used to construct models. This emergence of logical properties has the advantage of accounting for a puzzling phenomenon—the vagaries of everyday spatial relations. The inferences modelled in the program are for the *deictic* interpretation of 'on the right of', i.e. the relation as perceived from a speaker's point of view. However, some entities (e.g. human beings) have an intrinsic right-hand side and left-hand side (Miller and Johnson-Laird 1976, section 6.1.3). Hence the premises

The President is on Pat's right
Pat is on the Queen's right

can refer to the position of three individuals in relation to their intrinsic right-hand sides, which are independent of the speaker's point of view. A model of these spatial relations calls for the system to locate Pat, to establish a frame of reference around her based on her orientation, and then to use the semantics of 'on X's right' to add the President to the model in a position on the right-hand side of the lateral plane passing through Pat. The same semantics as the program uses for 'on the right' can be used, but instead of applying to the axes of the spatial array it applies to axes centred on individuals according to their orientations. If the individuals are seated in a line along one side of table, the model of them yields the transitive conclusion

The President is on the Queen's right (though not next to her).

However, if the individuals are seated round a small circular table, each premise can be true, but the transitive conclusion can be false: the President is opposite the Queen. Depending on the size of the table and the number of individuals seated around it, transitivity can occur over limited regions, and the meaning of the spatial relation and the mental model of the reference situation account for the variability in its logical properties.

The psychological evidence shows that individuals construct models in a broadly similar way to the program, both for spatial relations (Byrne and Johnson-Laird 1989) and for temporal relations (Schaeken *et al.* 1996). The results establish three main phenomena. First, when a description is consistent with just one model, reasoning is simple and participants typically draw over 90% correct conclusions. However, when a description is consistent with more than one model, there is a reliable decline in performance. Secondly, when participants draw their own conclusions, errors tend to arise because they fail to consider all the models of the premises. Thirdly, participants take reliably longer to read a premise that leads to multiple models than to read a corresponding premise in a one-model problem. However, some individuals take so little extra time that, as Schaeken (personal communication) has suggested, they may not be constructing multiple models. Instead, they may construct a single model that contains a referent represented as having one of several possible positions (Vandierendonck *et al.* 2000).

Readers should now understand the essential characteristics of mental models. A key feature of models is their structure. Hence a model of spatial or temporal relations is organized in terms of axes so that information in the model can be accessed by values on these axes. Such an organization in a mental model does not necessitate that information is laid out in a physical array in the brain. The spatial reasoning program relies on arrays, which are data structures that function as arrays rather than corresponding to physical arrays in the computer's memory.

4. Reasoning with sentential connectives

The model theory extends naturally to reasoning with negation and sentential connectives, such as 'if', 'or', and 'and'. The extension depends on assumptions about three main levels of human reasoning. They concern, first, the core meanings of sentential connectives, secondly, their modulation as a result of the meanings and referents of the clauses they connect, and, thirdly, the strategies that human reasoners rely on for sentential inference. No theory of human reasoning can afford to neglect any of these levels, and this section of the paper describes each of them.

4.1 *The core meanings of sentential connectives*

In sentential logic, connectives have *truth-functional* meanings, i.e. the truth-values of sentences formed with them depend solely on the truth-values of the clauses that they connect. Logicians capture these meanings in truth tables. Naive individuals, who know no logic, do not use truth tables. The size of a truth table doubles with each additional atomic proposition, but, as Osherson (1974–1976) noted, the difficulty of inferences does not double in this way. However, *fully explicit* models of possibilities are a step toward psychological plausibility. The fully explicit models of an exclusive disjunction of the form, *A or else B but not both*, are shown here in separate horizontal rows:

$$\begin{array}{ll} A & \neg B \\ \neg A & B \end{array}$$

where ¬ denotes negation, and A and B denote models of their respective propositions. Each model represents a different possibility compatible with the disjunction. Table 1 presents the fully explicit models for the logical connectives. Fully explicit models correspond to the true rows in the truth table for each connective, but they are more parsimonious than truth tables because they work on the assumption that true possibilities are exhaustively represented, and so there is no need to stipulate what is false.

Fully explicit models make possible an algorithm for sentential reasoning that has advantages over both standard proof theory and truth tables (Jeffrey 1991). The algorithm forms the Cartesian product of the models of the premises, i.e. it forms conjunctions of each model representing the current premise and each model representing all the previous premises. There are two rules for forming a conjunction of a pair of models.

Rule 1 If one model contains the negation of a proposition in the other model, then the result is the null model.

The null model is akin to the empty set, i.e. it is the model of a self-contradiction. For example, the conjunction of the model

$$A \quad \neg B$$

Table 1. The fully explicit models and the mental models of the possibilities compatible with sentences containing the logical connectives.

Sentences	Fully explicit models		Mental models	
A and B	A	B	A	B
Neither A nor B	¬ A	¬ B	¬ A	¬ B
A or else B but not both	A	¬ B	A	
	¬ A	B		B
A or B or both	A	¬ B	A	
	¬ A	B		B
	A	B	A	B
If A then B	A	B	A	B
	¬ A	B	. . .	
	¬ A	¬ B	. . .	
If, and only if A, then B	A	B	A	B
	¬ A	¬ B	. . .	

The symbol ¬ denotes negation, and the symbol . . . denotes a wholly implicit model. Each line represents a separate model.

with the model

¬ A

yields the null model.

Rule 2 Otherwise, the conjunction of a pair of models yields a single model of each proposition in the pair.

For example, the conjunction of the model

¬ A B

with the model

¬ A C

yields the model

¬ A B C

These two rules are also used recursively to construct the models of compound premises containing multiple connectives. Any connective can, of course, be expressed in terms of negation and conjunction, and the negation of a set of models is merely its complement from the set of all possible models based on the same atomic propositions. For example, the negation of the set

A ¬ B
¬ A B

is

A B
¬ A ¬ B

The two rules above allow any set of premises to be represented in a single set of models. For example, the premises

(1) A or else B but not both
(2) B or else C, but not both

have the models

(Premise 1)		(Premise 2)	
A	¬B	B	¬C
¬A	B	¬B	C

Their conjunction according to the two rules is

A	¬B	C
¬A	B	¬C

where the null models are not shown.

An inference is valid if its conclusion holds in all the possibilities that satisfy the premises. The algorithm is guaranteed to construct models of all these possibilities, and so any conclusion that holds in all the models of the premises is valid. The algorithm has two additional procedures. One procedure evaluates *given* conclusions. It shows that a conclusion is logically necessary if it holds in all the models, possible if it holds in some of the models, or impossible if it holds in none of the models. The other procedure finds a member of the set of most parsimonious descriptions of models (Johnson-Laird and Byrne 1991, Chapter 9). For example, given the set of models

A	B	¬C	D
A	B	C	¬D
A	B	¬C	¬D
¬A	B	C	D
A	¬B	C	D
¬A	¬B	C	D

it yields the parsimonious description

A and B, or else C and D.

The procedure is more efficient than the standard ‘prime implicant’ method. Therefore the system has the advantage that, unlike standard systems of proof, it can draw its own conclusions from premises. Likewise, it is more parsimonious than the use of truth tables although, like them, it is computationally intractable, i.e. NP-hard (Cook 1971).

Human reasoners do not normally use fully explicit models. They use *mental* models, which are more parsimonious because they follow the principle of truth (see section 2). Unlike a fully explicit model, a mental model represents only those atomic propositions in the premises that are true in a possibility. For example, the mental models of an exclusive disjunction *Not-A or else B* are as follows:

¬A	
	B

The first mental model does not represent B, which is false in this possibility, and the second mental model does not represent not-A, which is false in this possibility, i.e. A is true. The theory postulates that individuals make a mental footnote about what is false in a model, but they soon forget these footnotes even if they think of them in the first place. Hence they tend to neglect what is false in a possibility. However, if they do retain the footnotes, they are able to flesh out models explicitly. Another constraint on human performance is that individuals normally have just a single mental model in mind at a time.

The mental models of a conditional *If A then B* are

```
A        B
     ...
```

where the ellipsis denotes a model of the possibilities in which the antecedent of the conditional is false. The model has no explicit content, but allows for possibilities in which A is false. Individuals tend not to think explicitly about what holds in these possibilities. However, if they retain the footnote about what is false, then again they can flesh out these mental models into fully explicit models. The mental models of the biconditional *If, and only if, A then B* are identical to those for the conditional. What differs is that the footnote on the implicit model is that both A and B are false in this possibility. Table 1 presents the mental models based on the main sentential connectives.

The algorithm for reasoning with mental models contains additional rules to deal with implicit information. Table 2 summarizes the full set of rules for reasoning at the most primitive level, i.e. without taking mental footnotes into account. A program implemented by the author performs at four different levels of expertise depending on the use, if any, of mental footnotes. It ranges from the most primitive level (table 2) through to the use of fully explicit models according to the two rules given above.

The principle of truth predicts that reasoners should succumb to *illusory* inferences, which are compelling but invalid. This prediction is not an obvious consequence of the principle, but it was discovered in the output of the computer program. A simple example of an illusion is the following problem:

Only one of the following assertions 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 ten, or both.

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

Ninety-nine per cent of the undergraduates who tackled this problem responded ‘Yes’ (Goldvarg and Johnson-Laird 2000). They grasped that the first assertion allows two possibilities in which an ace occurs, and so they inferred that an ace is possible. However, if there were an ace in the hand, then both of the first two assertions would then be true, contrary to the rubric that only one of them is true. When the premises were paired with the question

Is it possible that there is a jack?

nearly all the participants again responded ‘Yes’. They considered that the third assertion, and its mental models, showed that there could be a jack. This time they

Table 2. The rules for updating existing mental models with the models of the current premise.

Rule 1	The conjunction of a pair of models representing respectively a proposition and its negation yield the null model, e.g. A ¬ B and ¬ A yield nil
Rule 2	The conjunction of a pair of models in which a proposition, such as B, in one model is not represented in the other model depends on the set of models of which this other model is a member. If B occurs in at least one of these models, then its absence in the current model is treated as negation, e.g. A B and A yields nil However, if B does not occur in one of these models (e.g. only its negation occurs in them) then its absence is treated as equivalent to its affirmation, and the conjunction follows the next rule
Rule 3	The conjunction of a pair of fully explicit models free from contradiction updates the second model with all the new propositions from the first model, e.g. A B and A yields A B
Rule 4	The conjunction of a pair of implicit models yields the implicit model: . . . and . . . yield . . .
Rule 5	The conjunction of an implicit model with a model representing propositions by default yields the null model, e.g. . . . and B C yield nil However, if none of the propositions (B C) are represented in the set of models containing the implicit model, then the conjunction yields the model of the propositions, e.g. . . . and B C yield B C

The rules make a conjunction of each model in the existing set with each model of the premise. They are for the most primitive level of performance, which ignores mental footnotes, but rules 1 and 3 also account for fully explicit models.

were correct; the inference is valid. Hence the focus on truth does not always lead to error, and experiments accordingly compare illusions with matching control problems for which the neglect of falsity should not affect accuracy (Johnson-Laird *et al.* 2000). The reader may wonder why we used questions of the form 'Is it possible that . . .?', which has many alternative interpretations in logic. The answer is that our participants understand the question much more readily than questions about consistency, and, as the model theory proposes, they respond 'Yes' if the proposition in the question holds in at least one mental model of a premise.

The computer program predicts that illusory inferences should be sparse in the set of all possible inferences. Nevertheless, experiments have corroborated their occurrence in reasoning based on sentential connectives, quantifiers, probabilities, causal relations, and deontic relations. Yang taught participants to think explicitly about what is true and what is false. The difference between illusions and control problems vanished, but performance on the control problems fell from almost 100% correct to about 75% correct (Yang and Johnson-Laird 2000). The principle of truth limits understanding, but it does so without participants realizing it. They are confident in their responses, whether they are correct or incorrect (Johnson-Laird and Savary 1999).

Sceptics tend to retort that the investigators are prey to the illusions rather than their participants, and that logic is not an appropriate criterion to assess human deductive reasoning. However, the prediction of the illusions depends, not on logic, but on how individuals interpret the connectives in other simple assertions, i.e. the possibilities that they list for simple conjunctions, disjunctions, and conditionals. Moreover, if you think that the illusory responses are correct, then how do you explain the effects of the remedial procedures? You should argue that *they* yield illusions. Critics also argue that the illusions occur because the problems are artificial. In fact, a search of the Internet yields many illusions (Johnson-Laird and Savary 1999). And the alleged artificiality of the problems does not prevent the participants from responding correctly to the control problems. Human reasoners have limited working memories, and they conserve processing capacity by abiding by the principle of truth. Most of the time the neglect of falsity is innocuous. Just occasionally, however, it yields models of possibilities that are incompatible with the meanings of assertions. These possibilities, in turn, yield systematic and predictable illusions.

Sentential reasoning is an excellent domain in which to rebut a complaint against the model theory. Critics argue that a major, if not fatal, problem with mental models is that they do not yield step-by-step arguments, and that to justify correct answers to the sorts of inferences that I have described calls for a formally valid argument. This criticism is mistaken, because the model theory does yield step-by-step arguments. For example, consider the following inference (from Van der Henst *et al.* 2002).

There is a red marble if and only if there's a green marble.
Either there is a green marble or else a blue marble, but not both.
There is a blue marble if and only if there is a white marble.
Does it follow that if there is a red marble then there is not a white marble?

The first premise is consistent with two possibilities, shown here as fully explicit models:

red	green
¬ red	¬ green

The conjunction with the second premise updates the two possibilities:

red	green	¬ blue
¬ red	¬ green	blue

Hence it follows that if there is a red marble, then there is not a blue marble. The conjunction with the third premise updates the possibilities:

red	green	¬ blue	¬ white
¬ red	¬ green	blue	white

Hence, it does follow validly that if there is a red marble, then there is not a white marble. This way of establishing validity is analogous to the use of a truth table, and it does not depend on the use of formal rules of inference. Naive individuals can not only understand such justifications, but they are also able to construct them.

4.2 *The modulation of core meanings*

The core meanings of the main sentential connectives were outlined in the previous section. These meanings appear to be truth functional (see table 1) but, in fact, no connectives in natural language are truth functional. A connective such as ‘before’ as in

The butler cleared his throat before he entered the room

is obviously not truth functional. The mere truth of each of its clauses does not suffice to establish that the sentence is true. The butler could have entered the room and then cleared his throat. According to the model theory, all sentential connectives including those in table 1 have core interpretations that can be modulated by several separate factors: the meanings of the clauses that they interconnect, co-referential relations between these clauses, the context in which the clauses are used, and general knowledge (Johnson-Laird and Byrne 2002). Modulation has three effects: it can add information to mental models, prevent their construction, and flesh them out into fully explicit models. The model theory postulates that the underlying mechanism is that knowledge, which is represented in fully explicit possibilities, modulates the mental models of assertions. In the case of inconsistency, knowledge takes precedence by default over the core interpretation of a connective.

As an illustration of modulation by the meaning of clauses, consider the disjunction:
Either it rained or it poured.

Its fully explicit models (see table 1), if they were unconstrained by co-reference and meaning, would be:

rain	¬ pour
¬ rain	pour
rain	pour

However, the meaning of *pour* entails that it rained, and therefore blocks the construction of the second model. The disjunction accordingly asserts that it rained and that it may have poured. Modulation, in turn, affects the inferences that reasoners draw from premises (Johnson-Laird and Byrne 2002).

The author has implemented a computer program that models the pragmatic modulation of assertions by knowledge. Its effects are illustrated in an old philosophical conundrum:

If the match was struck properly then it lit.
The match was soaking wet and it was struck properly.
What follows?

According to logic and the fully explicit models in table 1, it follows that the match lit. Human reasoners do not draw this conclusion, and nor does the program. Knowledge that wet matches do not light overrides the inference. The program starts by constructing the mental model of the premises:

match wet   match struck   match lights

If a match is soaking wet it does not light, and the program has a knowledge base containing this information in fully explicit models of the three possibilities:

match wet	¬ match lights
¬ match wet	¬ match lights
¬ match wet	match lights

The second premise states that the match is wet, which triggers the matching possibility in the preceding models:

match wet ¬ match lights

The conjunction of this model with the model of the premises would yield a contradiction, but the program, following the principle of modulation, gives precedence to knowledge and so yields the model

match wet match struck ¬ match lights

and so the match does not light. The model of the premises contains *match lights*, which triggers another possibility from the knowledge base above:

¬ match wet match lights

This possibility and the model of the premises are used to construct a counterfactual conditional.

> If it had not been the case that *match wet* and given *match struck* then it should have been the case that *match lights*.

Modulation is rapid and automatic, and it affects comprehension and reasoning (Newstead *et al.* 1997, Johnson-Laird and Byrne 2002).

If a sentence contains a truth-functional connective, then it is possible to assign it a truth value solely from the truth values of its constituent clauses. In contrast, the connectives in natural language are not truth-functional. Even those connectives in table 1 are not truth-functional, because it is always necessary to check whether their content and context modulate their interpretation. The interpretive system accordingly needs access to the meaning and reference of constituent clauses, not just their truth values.

4.3 *Strategies in reasoning*

When logically naive individuals are tested in experiments on reasoning, they begin with only rough ideas of how to proceed. They can reason, but not efficiently. As they gain experience, they spontaneously develop different strategies. They do so even without any feedback about the accuracy of their reasoning. Deduction itself may be a strategy (Evans 2000), and individuals may use it more in the West than the East (Nisbett 2003). However, the popularity of deductive reasoning puzzles, the so-called 'Su Doku' puzzles, in both Japan and the UK suggests that individuals in both cultures can be highly competent reasoners. Individuals also develop different strategies within deductive reasoning, and this section illustrates strategies in two main domains: reasoning with quantifiers and reasoning with sentential connectives.

The model theory began with an account of reasoning with quantifiers (Johnson-Laird 1983), as in syllogisms such as:

Some of the artists are beekeepers.
All the beekeepers are chemists.
Therefore, some of the artists are chemists.

The theory postulated that individuals construct models of the possibilities compatible with the premises and draw whatever conclusion, if any, holds in them. The preceding inference is easy. Yet each premise is compatible with several possibilities, and together they are compatible with 16 distinct possibilities. Thus the second premise is consistent with the existence, or not, of chemists who are not beekeepers, and with the existence, or not, of individuals who are neither beekeepers nor chemists. However, the inference can be made without explicitly representing all these possibilities. The program implementing the theory accordingly constructs this model of the premises:

```
artist   [beekeeper]   chemist
artist
         [beekeeper]   chemist
              . . .
```

where each row represents a different sort of individual, the ellipsis represents the possibility of other sorts of individual, and the square brackets indicate that beekeepers have been represented exhaustively, i.e. no more can be added to the model in fleshing out the ellipsis. The model yields the conclusion

Some of the artists are chemists.

or its converse

Some of the chemists are artists.

There are many ways in which reasoners might use such models, and the model theory postulates that reasoners develop a variety of strategies from exploring manipulations of models (Bucciarelli and Johnson-Laird 1999). A basic strategy has been modelled in a computer program which searches for counter-examples. What follows is a typical output from the program. The first line shows the two premises, and then the initial model. The program scans the model in both directions to draw two conclusions. It then finds an alternative model, and revises its conclusions accordingly. Finally, it finds a further model that is a counter-example to one of the two putative conclusions.

```
NO A ARE B.   ALL B ARE C.
[A]  –B
[A]  –B
        [B]  C
        [B]  C
```

NO A ARE C.   NO C ARE A.

[A]   −B   C

[A]   −B

        [B]   C

        [B]   C

SOME A ARE NOT C.   SOME C ARE NOT A.

[A]   −B   C

[A]   −B   C

        [B]   C

        [B]   C

NO VALID CONCLUSION.   SOME C ARE NOT A.

The counter-example eliminates the conclusion, *Some A are not C*, and so only the other conclusion survives unscathed. In other cases, of course, counter-examples refute all categorical conclusions interrelating the end terms of the syllogism.

Stenning and Yule (1997) have argued that individuals organize syllogistic reasoning as a search for individuals who necessarily exist given the premises; for example, given the premise *Some A are B*, there must be an A who is B. They proposed three different algorithms that all yield the same valid syllogistic inferences. One algorithm uses Euler circles supplemented with a notation for necessary individuals, one uses representations of individuals in line with the model theory, and one uses verbal rules, such as 'If there are two existential premises, i.e. that contain "some", then respond that there is no valid conclusion'. Stenning and Yule concluded that reasoners may develop different representational systems depending on the task. However, their three systems yield only correct responses. The difficulty is to explain the systematic errors that individuals make if they are using formal rules of inference. If the rules are valid, they cannot yield systematic errors. However, if they include invalid rules, then the system is likely to be internally inconsistent. In contrast, the model theory accounts for systematic errors in terms of the principle of truth and fails to consider all possible models.

The *external* models that reasoners constructed with cut-out shapes corroborated the hypothesis that individuals develop various strategies (Bucciarelli and Johnson-Laird 1999). They may focus on necessary individuals, as Stenning and Yule showed in their experimental task, but the typical representations of premises include individuals who are not necessary. For example, the typical representation of *Some of the A are B* is

A   B

A   B

A

Hence, the focus on necessary individuals is a particular strategy, but its implementation may require the representation of other individuals. For example, a representation of necessary individuals alone does not account for this sort of invalid inference, which many individuals make:

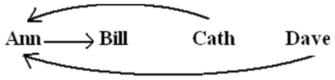
Some A are B.
Some B are C.
Therefore some A are C.

The model theory has been extended to inferences based on premises containing more than one quantifier. Consider, for example, this inference from a recent study (Cherubini and Johnson-Laird 2004):

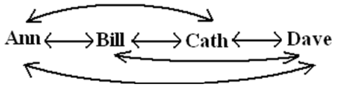
There are four persons: Ann, Bill, Cath, and Dave.
Everybody loves anyone who loves someone.
Ann loves Bill.
What follows?

Most people can construct model 1 in which arrows denote the relation of *loving*. Hence they infer that everyone loves Ann. However, if you ask them whether it follows that Cath loves Dave, they usually respond 'No'. They are mistaken, but the inference calls for a recursive use of the quantified premise. It is applied to model 1 to yield in turn model 2 (strictly speaking, all four persons love themselves too). Hence, Cath does loves Dave, and individuals are able to grasp the validity of the inference if it is explained to them with diagrams.

Model 1



Model 2



Even in simple sentential reasoning, individuals develop different strategies (Van der Henst *et al.* 2002). Consider again the earlier problem of the following form, in which each clause refers to a different coloured marble in a box:

A if and only if B.
Either B or else C, but not both.
C if and only if D.
Does it follow that if A then not D?

When logically naive individuals first encounter such a problem, they flail around for a while before they are able to figure out the correct response, which they nearly all make. With experience of this sort of problem but no feedback about accuracy, different individuals develop different strategies. Some people develop a strategy based on suppositions (Byrne and Handley 1997). When they think aloud about the preceding problem, they say, for example:

> *Suppose A*. It follows from the first premise that B. It follows from the second premise that not C. The third premise then implies not D. So, yes, the conclusion follows.

Some individuals instead construct a chain of conditionals leading from one clause in the conclusion to the other, for example *If A then B, If B then not C, If not C then* *not D*. Others develop a strategy in which they enumerate the different possibilities compatible with the premises. For example, they draw a vertical line down the page and write down the two possibilities that the premises yield:

A	
B	C
	D

Victoria Bell (unpublished studies) taught individuals this strategy, and she compared their performance with control participants and with participants whom she taught to use suppositions. Those who were taught to enumerate possibilities were the fastest and most accurate in drawing conclusions.

The nature of the premises and the conclusion can bias reasoners to adopt a predictable strategy; for example, conditional premises encourage the use of suppositions, whereas disjunctive premises encourage the enumeration of possibilities (Van der Henst *et al*. 2002). In sentential reasoning, there is a robust result regardless of which strategies individuals develop: inferences that call for only one mental model are easier than those that call for more than one model (see also Espino *et al*. 2000). Different strategies could reflect different mental representations, as Stenning and Yule (1997) suggest, but those strategies so far discovered are all compatible with the use of mental models. However, given enough experience of a class of problems, individuals do begin to notice formal patterns.

5. How diagrams can aid reasoning

The late Herbert Simon (1991, p. 96) described a nice piece of heterogeneous reasoning in his autobiography. Early in his career, he used circuit diagrams to help engineers to understand Supreme Court cases: the switch positions corresponded to the yes–no decisions of the Court. Later, Simon and his colleagues distinguished the role of diagrams in three separate sorts of process: search, recognition, and inference (Larkin and Simon 1987; Tabachneck and Simon 1992). Larkin and Simon claimed that diagrams can make it easier to find relevant information: you can scan from one element to another element nearby much more rapidly than you can find the same information in a set of assertions. They also claimed that diagrams can make it easier to identify instances of a concept, and that symmetries in a diagram can reduce the number of cases that need to be examined. However, about inference, they wrote:

> In view of the dramatic effects that alternative representations may produce on search and recognition processes, it may seem surprising that the differential effects on inference appear less strong. Inference is largely independent of representation if the information content of the two sets of inference rules [one operating on diagrams and the other operating on verbal statements] is equivalent—i.e. the two sets are isomorphs as they are in our examples. (Larkin and Simon 1987, p. 71)

Logicians have also sometimes claimed that diagrammatic methods of reasoning are improper (Tennant 1986), but the late Jon Barwise and his colleagues have shown that such methods can be valid, and indeed yield complete systems that capture all valid inferences (Barwise and Etchemendy 1992; Shin 1992). Barwise and Etchemendy (1994) developed a computer program, Hyperproof, that helps users to learn logic. It exploits the heterogeneity of reasoning, using diagrams to represent conjunctions but sentences to represent disjunctions:

> ...diagrams and pictures are extremely good at presenting a wealth of specific, conjunctive information. It is much harder to use them to present indefinite information, negative information, or disjunctive information. For these, sentences are often better. (Barwise and Etchemendy 1992, p. 82)

Can diagrams help the process of reasoning? Although all the authors above are pessimistic, their theories do not imply that diagrams can never aid inference. Instead, researchers have been unable to think of how diagrams might help, especially with reasoning depending on disjunctions and negations. Yet Peirce wrote of his 'existential' diagrams, which capture the whole of the predicate calculus, that they '[render] literally visible before one's very eyes the operation of thinking *in actu*' (Pierce 1931–1958, Vol. 4, para. 6), and 'put before us moving pictures of thought' (Vol. 4, par. 8).

The model theory predicts that diagrams can improve reasoning. A critical problem in reasoning is to keep track of the disjunctive possibilities compatible with the premises. It follows that a diagram that makes the possibilities explicit should help individuals to reason. The perception of such a diagram yields models (Marr 1982), and vision is a more direct route to models than the comprehension of a description. As we saw earlier, the sentences in a description need to be parsed with a compositional semantics to construct the representation of their meanings, which in turn are used to construct models. It follows that it should be easier to reason from diagrams that make possibilities explicit than from verbal descriptions of the same possibilities.

Bauer and Johnson-Laird (1993) tested this prediction experimentally. We used deductive problems of the following type, which are known as 'double disjunctions'.

Julia is in Atlanta or Raphael is in Tacoma, but not both.
Julia is in Seattle or Paul is in Philadelphia, but not both.
What follows?

Each of these exclusive disjunctions has two mental models, and their combination yields the following three mental models, which each represent a different possibility:

((Julia) Atlanta)		((Paul) Philadelphia)
((Julia) Seattle)	((Raphael) Tacoma)	
	((Raphael) Tacoma)	((Paul) Philadelphia)

where ((Julia) Atlanta) denotes a model of Julia in Atlanta. A conclusion that follows from these models is:

Ralph is in Tacoma or Paul is in Philadelphia, or both.

The model theory predicts that such problems based on exclusive disjunctions should be easier than those based on inclusive disjunctions:

Julia is in Atlanta or Raphael is in Tacoma, or both.
Julia is in Seattle or Paul is in Philadelphia, or both.
What follows?

This problem yields five models:

((Julia) Atlanta)		((Paul) Philadelphia)
((Julia) Seattle)	((Raphael) Tacoma)	
	((Raphael) Tacoma)	((Paul) Philadelphia)
((Julia) Atlanta)	((Raphael) Tacoma)	((Paul) Philadelphia)
((Julia) Seattle)	((Raphael) Tacoma)	((Paul) Philadelphia)

However, as in the previous problem, these models also support the valid conclusion:

Raphael is in Tacoma or Paul is in Philadelphia, or both.

When the problems are presented verbally, the model theory's prediction is corroborated (Johnson-Laird and Byrne 1991). Ironically, many psychological theories based on formal rules of inference have no rules for exclusive disjunctions, but translate them into conjunctions: *p or q or both, and not both p and q* (e.g. Rips 1994; Braine and O'Brien 1998). Therefore the theories predict that an inference of the form

p or q
not p
Therefore q

should be *harder* with an exclusive disjunction than with an inclusive disjunction. In contrast, the evidence shows that inferences are easier from exclusive disjunctions than from inclusive disjunctions.

In one experiment, Johnson-Laird and Bauer compared a verbal presentation of double disjunctions with the use of diagrams of the type shown in figure 1. This diagram represents the spatial relations in an iconic way; for example, the word 'Julia' lies within an ellipse labelled 'Atlanta'. However, the representation of the alternative possibilities is not iconic; the box at the center of the two lines connecting the two ellipses is a conventional symbol representing inclusive disjunction. To represent an exclusive disjunction, there was instead a circle containing a cross. Both the percentages of accurate conclusions and their latencies showed that the exclusive disjunctions were easier than the inclusive disjunctions, but the diagrams (28% correct conclusions) did not reliably enhance performance in comparison with verbal premises (30% correct conclusions). Double disjunctions remained difficult and the diagrams were no help.

The diagram in figure 1 does not enable reasoners to envisage the alternative possibilities compatible with the premises. We devised a new sort of diagram, illustrated in figure 2, which is analogous to a simple electrical circuit. The participants' task was to complete a path from one side of the diagram to the other by moving the shapes corresponding to people into the slots corresponding to cities. In a second experiment, four separate groups of participants tackled logically

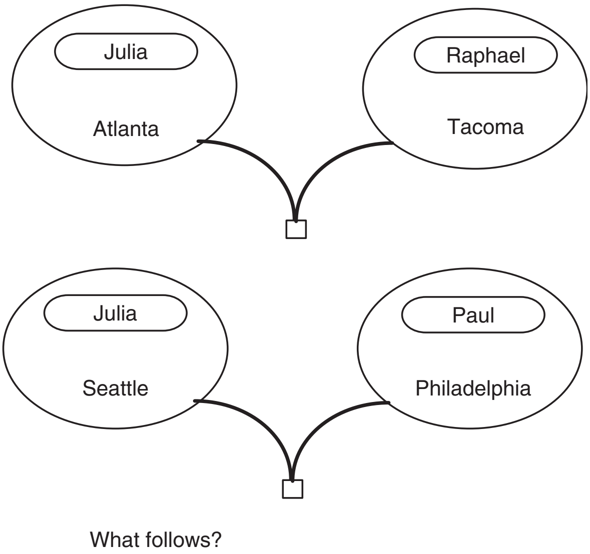


Figure 1. The diagram representing a double disjunction (an inclusive one) in the first diagram experiment.

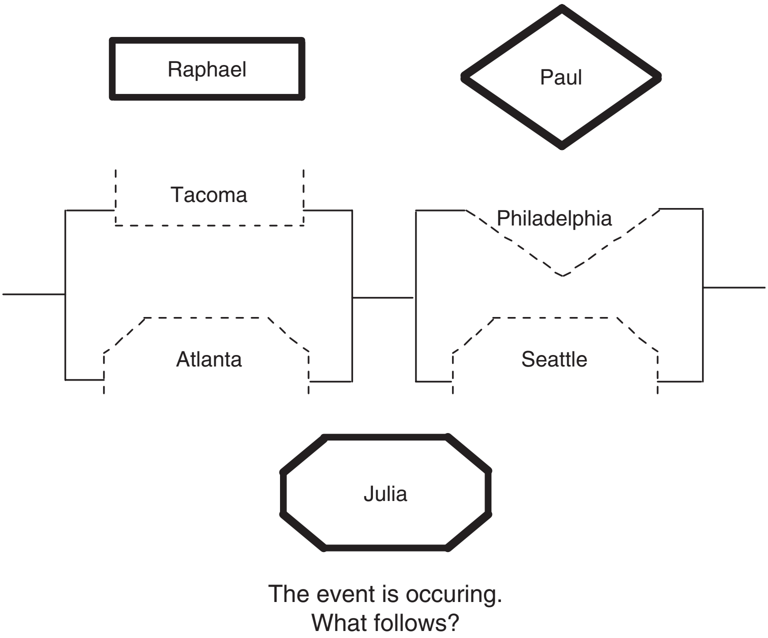


Figure 2. The diagram representing a double disjunction (an inclusive one) in the second diagram experiment.

equivalent problems. Two groups received problems about people in cities, which were presented in diagrams (e.g. figure 2) to one group and in verbal premises to the other group. Two further groups received problems about electrical switches, which were presented in circuit diagrams to one group and in verbal premises to the other group. The content of the problems, whether they were about people or switches, had no reliable effect on performance. The mode of presentation, whether the problems were in diagrams or verbal premises, had a striking effect. There were 74% correct responses to the diagrammatic problems compared with only 46% correct responses to the verbal problems. Once again, exclusive disjunctions were easier than inclusive disjunctions. The latencies of the responses had exactly the same pattern; for example, the participants responded much faster to the diagrammatic problems (a mean of 99 seconds) than to the verbal problems (a mean of 135 seconds). The participants tended to err by overlooking possibilities. They made a smaller proportion of errors that were inconsistent with the premises, and hardly any such errors with the diagrams.

A 30% improvement in the accuracy of reasoning and a speed up of 30 seconds are dramatic effects. The moral is that human reasoners do represent the possibilities compatible with premises, whether they are statements or diagrams. The diagrams in the second experiment make it easy to think about negation and about disjunctive possibilities. Individuals perceive the layout and in their mind's eye they can move people into places and out again, and they can readily enumerate the different possibilities. The simple ways to complete the circuit in figure 2 are to put Raphael in Tacoma and Paul in Philadelphia, or Raphael in Tacoma and Julia in Seattle, or Paul in Philadelphia and Julia in Atlanta. A listing of this sort, in 'disjunctive normal form', is what most individuals construct, although they may overlook a possibility. They seem to appreciate that in the first case, in which Raphael and Paul are in their respective cities, it does not matter where Julia is. Likewise, some reasoners grasp that it follows that in all the possibilities Raphael is in Tacoma or Paul is in Philadelphia or both. The diagram accordingly makes available to them the crucial aspects of the possibilities in a more salient way than the verbal premises do. They manipulate the mental model underlying the visual image, and in this way construct the alternative possibilities more readily than they can from verbal premises. It follows that, contrary to a common view in cognitive science (e.g. Pylyshyn 1973; Palmer 1975), diagrams are not merely encoded in propositional representations equivalent to those constructed from verbal premises.

Are mental models merely images? Readers may be tempted to make this assumption, but they should resist the temptation. When individuals manipulate visual images, the evidence suggests that they are manipulating, not a two-dimensional 'picture-like' representation, but an underlying mental model of a three-dimensional entity (Shepard and Metzler 1971, Metzler and Shepard 1982, p. 45). Likewise, when individuals reason about matters that elicit vivid images, such as the relations 'cleaner than' and 'dirtier than', their reasoning is slowed down, and it recruits additional areas of the brain that mediate vision (Knauff and Johnson-Laird 2002; Knauff *et al.* 2003). Accordingly, visual images differ from models. Images represent visualizable entities, properties, and relations from a particular point of view. They are projected from the visualizable aspects of underlying models. In contrast, models can be three-dimensional and can embody abstract predicates.

Hence they can represent any situation, and operations on them can be purely conceptual.

6. Conclusions

This paper began with a typical example of heterogeneous reasoning from daily life. You inferred that you would get wet, and your inference was based on a visual observation that it was raining, from your knowledge that you would get wet without an umbrella, and from your spouse's remark that the umbrella was lost. All this information could be converted into expressions in a formal language, and the conclusion proved in a logical calculus. However, human reasoners do not appear to reason in this way. What seems central to their reasoning is a representation of what is possible. This intuition lies at the heart of the theory of mental models. Each mental model represents a possibility, and so it is feasible to base reasoning, both deductive and inductive, on the manipulation of models. In the case of reasoning about spatial relations, a model-based theory and its computer implementation have no need of meaning postulates to capture the logical properties of relations, such as transitivity, symmetry, and reflexivity. They emerge from the meanings of relational expressions and the machinery for manipulating models. One happy by-product of this system is that it captures the vagaries of logical properties. The transitivity of being on a person's right, for example, depends on the seating arrangement. It would be a major headache to try to deal with this problem by relying on meaning postulates.

The model theory extends naturally to reasoning based on sentential connectives. It has several advantages over accounts based on formal rules of inference. With fully explicit models, the representation of a set of premises captures all the possibilities compatible with the premises. Hence, on the one hand, any conclusion that holds in all of these models is deductively valid; on the other hand, an invalid conclusion can be refuted by the construction of a counter-example, i.e. a model that satisfies the premises but refutes the conclusion. Without recourse to counter-examples, the only method of establishing invalidity is to fail to find a proof of the conclusion. This procedure cannot explain how we can *know* that an inference is invalid (Barwise 1993). Sceptics sometimes suppose that counter-examples can be used only in a system based on formal rules of inference, but in logic they function perfectly well in truth-tables, where a counter-example corresponds to a row showing that a conditional from premises to conclusion is not a tautology. There is no need to embed counter-examples in formal proofs, and to deny this point is to make nonsense of the proofs of completeness and soundness of the sentential calculus. Likewise, programs such as the one described for syllogisms use counter-examples as an integral part of model-based reasoning. Unlike logic, however, the meaning of sentential connectives in daily life is modulated by the meaning of words and clauses, by co-referential relations, and by knowledge. The model theory shows how modulation works. One of the consequences of modulation is that sentential connectives in natural language are never truth-functional. Their interpretation, even though it may correspond to a truth-functional analysis in certain utterances, cannot be truth-functional because the interpretative system must always check whether modulation has altered the core meaning.

One unfortunate aspect of human reasoning arises from the principle of truth. Because human working memory is limited in capacity, human reasoners cannot rely on truth tables. Their mental models represent atomic propositions in the premises only when they are true in a possibility. The failure to represent what is false seems innocuous. Indeed, for several years, the present author was oblivious of its serious consequences. The output of a computer program implementing the theory revealed the occurrence of radical discrepancies between the possibilities represented in mental models and the possibilities represented in fully explicit models. The results of psychological experiments have shown that human reasoners rely on mental models rather than fully explicit models. Hence they succumb to systematic illusory inferences, such as the following example.

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

Most participants in a recent experiment said 'Yes' (Walsh and Johnson-Laird 2004). They envisaged just the two possibilities corresponding to the mental models of the disjunction. However, the fully explicit models show that if the second conjunction is true, then one way in which the first conjunction could be false is precisely if Jane is kneeling by the fire and *not* looking at the TV. The principle of truth, alas, leads to predictable fallacies.

Diagrams have a long pedagogical history in logic and in life. Many theorists have supposed that they can help us to recover important information more readily. Hence if you want to convey what is going on in a complex domain, such as the flow of air around an airfoil, then the translation of the data into a visual display can capitalize on the power of the visual system to extract high-level patterns from low-level data. To make sense of an array of 100 million numbers (the intensities of light falling on the cells in the retina) the brain has 'software' that uses these data to construct a high-level model of the world suitable for the limited powers of consciousness. The use of diagrams in logic has, with the glorious exception of Peirce's system, been as a supplement to rather than an intrinsic part of reasoning. Theorists have supposed that diagrams do not enhance the process of inference. The model theory takes a different view. Human reasoning is based on the manipulation of models of possibilities rather than the derivation of conclusions from expressions in a mental language. Hence certain diagrams should yield helpful models more directly than verbal premises do. The evidence corroborates this view. The difficulty of reasoning about disjunctive possibilities can be ameliorated with a diagram. One can imagine moving a shape from one position to another and in this way to envisage possibilities. Thus one is able to reason more rapidly and more accurately. The phenomena vindicate the theory of mental models because, unlike other accounts, it predicts them.

In sum, human reasoning is based on heterogeneous sources. It integrates these diverse forms of information in mental models of the corresponding possibilities. These models are as iconic as possible, but they contain many elements that are not visualizable. Their evolutionary origin may be in the viewer-independent spatial representations that the mammalian visual system constructs.

Acknowledgements

The author thanks Selmer Bringsjord, Yingrui Yang, and an anonymous reviewer for many helpful criticisms. This research was supported in part by a grant from the National Science Foundation (BCS 0076287) to investigate strategies in reasoning.

References

J. Barwise, "Everyday reasoning and logical inference", *Behav. Brain Sci.*, 16, pp. 337–338, 1993.

J. Barwise and J. Etchemendy, "Hyperproof: logical reasoning with diagrams", in *AAAI Symposium on Reasoning with Diagrammatic Representations*, N.H. Narayanan, Ed., Stanford, CA: Stanford University, 1992, pp. 80–84.

J. Barwise and J. Etchemendy, *Hyperproof*, Stanford, CA: CSLI Publications, 1994.

J. Barwise and J. Etchemendy, "Heterogeneous logic", in *Diagrammatic Reasoning: Cognitive and Computational Perspectives*, J. Glasgow, N.H. Narayanan and B. Chandrasekaran, Eds., Cambridge, MA: MIT Press, 1995, pp. 211–234.

M.I. Bauer and P.N. Johnson-Laird, "How diagrams can improve reasoning", *Psychological Science*, 4, pp. 372–378, 1993.

M.D.S. Braine and D.P. O'Brien (Eds), *Mental Logic*, Mahwah, NJ: Erlbaum, 1998.

M. Bucciarelli and P.N. Johnson-Laird, "Strategies in syllogistic reasoning", *Cogn. Sci.*, 23, pp. 247–303, 1999.

R.M.J. Byrne and S.J. Handley, "Reasoning strategies for suppositional deductions", *Cognition*, 62, pp. 1–49, 1997.

R.M.J. Byrne and P.N. Johnson-Laird, "Spatial reasoning", *J. Mem. Lang.*, 28, pp. 564–575, 1989.

P. Cherubini and P.N. Johnson-Laird, "Does everyone love everyone? The psychology of iterative reasoning", *Think. Reasoning*, 10, pp. 31–53, 2004.

S.A. Cook, "The complexity of theorem proving procedures", in *Proceedings of the Third Annual Association of Computing Machinery Symposium on the Theory of Computing*, 1971, pp. 151–158.

K. Craik, *The Nature of Explanation*, Cambridge: Cambridge University Press, 1943.

O. Espino, C. Santamaría, E. Meseguer and M. Carreiras, "Eye movements during syllogistic reasoning", in *Mental Models in Reasoning*, J.A. García-Madruga, N. Carriedo and M.J. González-Labra, Eds, Madrid: Universidad Nacional de Educación a Distancia, 2000, pp. 179–188.

J.St.B.T. Evans, What could and could not be a strategy in reasoning, in *Deductive Reasoning and Strategies*, W.S. Schaeken, G. De Vooght, A. Vandierendonck and G. d'Ydewalle, Eds., Mahwah, NJ: Erlbaum, 2000, pp. 1–22.

Y. Goldvarg and P.N. Johnson-Laird, "Illusions in modal reasoning", *Mem. Cognition*, 28, pp. 282–294, 2000.

J.Y. Halpern and M.Y. Vardi, "Model checking vs. theorem proving: a manifesto", in *Proceedings of the International Conference on Principles of Knowledge Representation and Reasoning*, J.A. Allen, R. Fikes, and E. Sandewall, Eds, San Mateo, CA: Morgan Kaufmann, 1991, pp. 325–334.

B. Inhelder and J. Piaget, *The Growth of Logical Thinking from Childhood to Adolescence*, London: Routledge & Kegan Paul, 1958.

R.C. Jeffrey, *Formal Logic, its Scope and Limits*, 3rd ed., New York: McGraw-Hill, 1991.

P.N. Johnson-Laird, *Mental Models: Towards a Cognitive Science of Language, Inference, and Consciousness*, Cambridge, MA: Harvard University Press, 1983.

P.N. Johnson-Laird, "Mental models and deduction", *Trends Cogn. Sci.*, 5, pp. 434–442, 2001.

P.N. Johnson-Laird, The history of mental models, in *Psychology of Reasoning: Theoretical and Historical Perspectives*, K. Manktelow and M.C. Chung, Eds., Hove, Sussex: Psychology Press, 2004, pp. 179–212.

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

P.N. Johnson-Laird and R.M.J. Byrne, "Conditionals: a theory of meaning, pragmatics, and inference", *Psychol. Rev.*, 109, pp. 646–678, 2002.

P.N. Johnson-Laird and U. Hasson, "Counterexamples in sentential reasoning", *Mem. Cognition*, 31, pp. 1105–1113, 2003.

P.N. Johnson-Laird, P. Legrenzi, V. Girotto, M. Legrenzi and J.-P. Caverni, "Naive probability", *Psychol. Rev.*, 106, pp. 62–88, 1999.

P.N. Johnson-Laird, P. Legrenzi, P. Girotto and M.S. Legrenzi, "Illusions in reasoning about consistency", *Science*, 288, pp. 531–532, 2000.

P.N. Johnson-Laird and F. Savary, "Illusory inferences: a novel class of erroneous deduction", *Cognition*, 71, pp. 191–229, 1999.

M. Knauff and P.N. Johnson-Laird, "Imagery can impede inference", *Mem. Cognition*, 30, pp. 363–371, 2002.

M. Knauff, T. Fangmeier, C.C. Ruff and P.N. Johnson-Laird, "Reasoning, models, and images: behavioral measures and cortical activity", *J. Cogn. Neurosci.*, 4, pp. 559–573, 2003.

J. Larkin and H. Simon, "Why a diagram is (sometimes) worth 10,000 words", *Cogn. Sci.*, 11, pp. 65–99, 1987.

D. Marr, *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*, San Francisco, CA: W.H. Freeman, 1982.

J. Metzler and R.N. Shepard, "Transformational studies of the internal representations of three-dimensional objects", in *Mental Images and Their Transformations*, R.N. Shepard and L.A. Cooper, Eds, Cambridge, MA: MIT Press, 1982, pp. 25–71 (originally published in *Theories in Cognitive Psychology*: *The Loyola Symposium*, R.L. Solso, Ed., Hillsdale, NJ: Erlbaum, 1974).

G.A. Miller and P.N. Johnson-Laird, *Language and Perception*, Cambridge, MA: Harvard University Press, 1976.

S.E. Newstead, M.C. Ellis, J.St.B.T. Evans and I. Dennis, "Conditional reasoning with realistic material", *Think. Reasoning*, 3, pp. 49–76, 1997.

R.E. Nisbett, *The Geography of Thought: How Asians and Westerners Think Differently…and Why*, New York: Free Press, 2003.

D.N. Osherson, *Logical Abilities in Children*, Vols 1–4, Hillsdale, NJ: Erlbaum, 1974–1976.

S.E. Palmer, "Visual perception and world knowledge: notes on a model of sensory–cognitive interaction", in *Explorations in Cognition*, D.A. Norman, D.E Rumelhart, and the LNR Research Group, Eds, San Francisco, CA: W.H. Freeman, 1975.

C.S. Peirce, in *Collected Papers of Charles Sanders Peirce*, Vols 1–9, C. Hartshorne, P. Weiss and A. Burks, Eds., Cambridge, MA: Harvard University Press, 1931–1958.

Z.W. Pylyshyn, "What the mind's eye tells the mind's brain: a critique of mental imagery", *Psychol. Bull.*, 80, pp. 1–24, 1973.

W.V.O. Quine *Methods of Logic*, Third edition, London: Routledge, 1974.

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

W.S. Schaeken, P.N. Johnson-Laird and G. d'Ydewalle, "Mental models and temporal reasoning", *Cognition*, 60, pp. 205–234, 1996.

R.N. Shepard and J. Metzler, "Mental rotation of three-dimensional objects", *Science*, 171, pp. 701–703, 1971.

S.-J. Shin, A semantic analysis of inference involving Venn diagrams, in *AAAI Symposium on Reasoning with Diagrammatic Representations*, N.H. Narayanan, Ed., Stanford, CA: Stanford University, 1992, pp. 85–90.

H.A. Simon, *Models of My Life*, New York: Basic Books, 1991.

E.E. Smith, C. Langston and R.E. Nisbett, "The case for rules in reasoning", *Cogn. Sci.*, 16, pp. 1–40, 1992.

K. Stenning and P. Yule, "Image and language in human reasoning: a syllogistic illustration", *Cogn. Psychol.*, 34, pp. 109–159, 1997.

H. Tabachneck and H. Simon, "Effect of mode of presentation on reasoning about economic markets", in *AAAI Spring Symposium on Reasoning with Diagrammatic Representations*, N.H. Narayanan, Ed., Stanford, CA: Stanford University, 1992, pp. 56–64.

N. Tennant, "The withering away of formal semantics", *Mind Lang.*, 1, pp. 302–318, 1986.

J.-B. Van der Henst, Y. Yang and P.N. Johnson-Laird, "Strategies in sentential reasoning", *Cogn. Sci.*, 26, pp. 425–468, 2002.

A. Vandierendonck, G. De Vooght, C. Desimpelaere and V. Dierckx, "Model construction and elaboration in spatial linear syllogisms", in *Deductive Reasoning and Strategies*, W.S. Schaeken, G. De Vooght, A. Vandierendonck and G. d'Ydewalle, Eds., Mahwah, NJ: Erlbaum, 2000, pp. 191–207.

C. Walsh and P.N. Johnson-Laird, "Co-reference and reasoning", *Mem. Cognition*, 32, pp. 96–106, 2004.

Y. Yang and P.N. Johnson-Laird, "How to eliminate illusions in quantified reasoning", *Mem. Cognition*, 28, pp. 1050–1059, 2000.