Mental models: a gentle guide for outsiders

by
P.N. Johnson-Laird, Vittorio Girotto, and Paolo Legrenzi

April 7th 1998

Authors' address
(PJ-L): Department of Psychology
Princeton University
Princeton
NJ 08540
USA
tel: +1 609 258 4432
fax: +1 609 258 1113
email: phil@clarity. princeton.edu

(V.G.): Centre de Recherche en Psychologie Cognitive (CREPCO)
Université de Provence and CNRS (UMR 6561)
29 Avenue Robert Schuman
13621 Aix-en-Provence Cedex 1, France
email: girotto@newsup.univ-mrs.fr

(P.L.) Istituto di Psicologia
Facolta di Lettere e Filosofia
Universita di Milano
Milano
Italy
email: legrenzi@imiucca.csi.unimi.it

"All our ideas and concepts are only internal pictures".

-- Ludwig Boltzmann (1899)

Introduction

Mental models are psychological representations of real, hypothetical, or imaginary situations. They were first postulated by the Scottish psychologist Kenneth Craik (1943), who wrote that the mind constructs "small-scale models" of reality that it uses to anticipate events, to reason, and to underlie explanation. Like pictures in Wittgenstein's (1922) "picture" theory of the meaning of language, mental models have a structure that corresponds to the structure of what they represent. They are accordingly akin to architects' models of buildings, to molecular biologists' models of complex molecules, and to physicists' diagrams of particle interactions. Since Craik's original insight, cognitive scientists have argued that the mind constructs mental models as a result of perception, imagination and knowledge, and the comprehension of discourse. They study how children develop such models, how to design artifacts and computer systems for which it is easy to acquire a model, how a model of one domain may serve as analogy for another domain, and how models engender thoughts, inferences, and feelings.

When people first hear about mental models, as the epigraph from Boltzmann suggests, they suppose that they are nothing more than mental pictures or images. In fact, models are a more general notion both because some models cannot be visualized, and because images depend on underlying models. Thus, Roger Shepard and his colleagues have shown that individuals can start with a picture of a three-dimensional object and then mentally rotate the object into a different orientation (Metzler and Shepard, 1982). The rate of rotation is about 60° per second, and it holds both for rotations of objects in the plane of the picture and for their rotations in depth. It follows that people are rotating, not a two-dimensional image of the picture, but an underlying model of the three-dimensional object. Models can also represent abstract notions, such as negation and ownership, which are impossible to visualize. Psychological experiments have corroborated the existence of such abstract elements, and they have shown that reasoning is unaffected by how easy it is to visualize the

premises. The operations that can be carried out on images correspond to visual transformations, whereas the operations that can be carried out on models, as we shall see, can correspond to conceptual processes.

To illustrate a mental model, consider the description of a simple spatial arrangement of objects, which begins with the assertion:

A triangle is on the right of a circle.

Its meaning might be represented by an expression in a mental "language", but the situation it describes can be represented by a mental model. A computer program that we implemented constructs a spatial model:

o ?

Its structure is isomorphic to the actual spatial relation between the two objects. The model captures what is common to any situation in which a triangle is on the right of a circle, but it represents nothing about their size, color, distance apart, or other such matters. Where relevant, the model can be updated to take into account information about such matters. The spatial description continues, say, with the assertion:

The circle is above a square

and so it calls for a model of the three-dimensional relations among the objects. Our computer program constructs such models. As it shows, one important function of models is to yield inferences about unstated spatial relations among the objects. Indeed, a major use for mental models is to subserve reasoning of various sorts.

Our plan in the rest of this paper is to start with how models can be used to reason, and to contrast them with the orthodox view that reasoning is based on a sort of mental logic. In order to try to decide between the two competing theories, we then examine a crucial, and unexpected, consequence of the model theory. It predicts that reasoners are, in effect, programmed to commit certain systematic fallacies. We describe the predictions and report some illustrative results that demonstrate the occurrence of these fallacies -- some of which seduce even the most expert of reasoners. The fallacies are accordingly a new sort of cognitive illusion. The model theory also explains how people infer the probability of an event from their knowledge of the different ways in which the event can occur. We report the occurrence of systematic fallacies in this domain too, which we exploit as a sign that people are relying on mental models. A related phenomenon is what we refer to as "focusing", that is, the tendency for people to consider only what is represented in their models of a situation. Focusing, as we show, occurs in decision making: people focus on what they have represented in their models of the options. Finally, we apply models to the relatively unexplored domain of meta-reasoning, which is reasoning about what other people have reasoned -- a topic that is highly pertinent to economic behavior.

Mental models and formal rules of inference

Reasoning is a key component of human thinking, but no-one knows for sure how people reason. If they have been taught logic, they might fall back on it in a self-conscious way. But, if they have not been taught logic, are they incapable of right reasoning? Of course not. Early psychological studies solicited introspections about reasoning -- a procedure that was not very revealing, because people are not aware of how they reason. During the last twenty years, however, psychologists have formulated various theories of reasoning. Some say that it depends on a memory for previous examples, or on rules that capture general knowledge. These accounts, however, do not extend to the full inferential competence that most of us can display. We can make deductions about matters of which we know nothing:

If a chord sequence is tonal, then it can be generated by a context-sensitive grammar.

The twelve-bar blues has a chord sequence that is tonal.

? The twelve-bar blues has a chord sequence that can be generated by a context-sensitive grammar.

Even if you know nothing about tonal chord sequences or context-sensitive grammars, you can appreciate that this inference (of a form known as modus ponens) is valid, that is, the conclusion must be true given that the premises are true. Its validity hinges, not on general knowledge, but on linguistic knowledge about such sentential connectives as "if ___ then ___". The real controversy among psychologists is about this sort of deductive reasoning. On one side, there are those, such as Rips

(1994), who claim that it depends on formal rules of inference akin to those of a logical calculus. On the other side, there are those, such as ourselves (see also Johnson-Laird and Byrne, 1991), who claim that it is a semantic process that depends on mental models akin to the models that logicians invoke in formulating the semantics of their calculi. The controversy has been fruitful -- it has led to improvements in experimental methodology and in the theories themselves -- and so we will briefly consider the two opposing approaches.

To concentrate our minds, we will describe a robust phenomenon in reasoning. Consider, again, the previous inference about the twelve-bar blues. Given the premises, nearly everyone draws the valid conclusion. The inference is easy. But, now consider the following problem:

If the test is to continue then the turbine must be rotating fast enough to generate emergency electricity.

The turbine is not rotating fast enough to generate emergency electricity.

What, if anything, follows?

It happened that there were some engineers who knew that the premises were true, and yet they went on with the test. Why they continued is a mystery, because the test was not only dangerous, but pointless. It led to the tragic disaster at Chernobyl (Medvedev, 1990). We suspect that the engineers failed to draw the appropriate conclusion that the test should not continue. For several years, the first author has given groups of engineering students at Princeton an inference of the same form (which is known as modus tollens) with an abstract content. A substantial minority invariably responds, "nothing follows". People do indeed make errors in reasoning, and the difference in difficulty between the two inferences is robust. Even people who get both inferences right take reliably longer to make the second inference.

What underlies the difference in difficulty between the two inferences? According to formal rule theories (e.g. Rips, 1994), the mind is equipped with a set of formal rules of inference that it uses tacitly to make inferences. One such rule directly matches the form of the modus ponens inference about the twelve-bar blues:

If A then B

A

? B

and so the inference is easy to make. Reasoners match the premises to the form of the rule, and draw the conclusion corresponding to B. But, according to these theories, the mind does not have a rule that matches the form of the Chernobyl inference:

If A then B

Not B

? Not A.

Instead, the inference depends on the following chain of deductions, where each step is sanctioned by a mental rule of inference:

A (a supposition made for the sake of argument)

? B (using the rule for modus ponens on the first premise and the supposition)

?? B and not B (using a rule for forming a conjunction -- of the previous assertion and a premise)

? not A (using the rule of reductio ad absurdum, which yields the negation of any supposition leading to a contradiction)

The longer derivation of the inference thus accounts for its greater difficulty. Presumably, reasoners fail to find the required

sequence of steps, and so assume that nothing follows from the premises.

Readers familiar with twentieth century logic will know that logicians draw a sharp distinction between formal, syntactic, methods (proof theory) and semantic methods (model theory). Formal rule theories in psychology are based on a version of proof theory in which many rules are used so that proofs are intuitively obvious (the method of "natural deduction"). The mental model theory of reasoning is likewise analogous to model theory in logic (its most familiar branch is the method of "truth tables").

The mental model theory assumes that logically-untrained reasoners are not equipped with formal rules of inference, but rather rely on their ability to understand premises. They build mental models of the relevant states of affairs based on this understanding and on general knowledge. They can formulate a conclusion that is true in these models. And they can test its validity by establishing that no alternative models of the premises refute it. In other words, a mental model is a representation of a possibility, which itself may occur in many ways, and so its structure and content capture what is common to these many ways. An "exclusive" disjunction, such as:

There is a king or there is an ace in the hand, but not both has two alternative models to represent the two possibilities:

King

Ace

where this diagram is based on the convention that each row denotes a separate model. Likewise, a conditional such as:

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

calls for one explicit model of the salient possibility (the presence of the king and the ace) and an implicit model that merely allows for other possibilities (in which the conditional's antecedent, there is a king, is false) without spelling them out explicitly:

King Ace

. . .

The ellipsis represents the implicit model of the possibilities in which the antecedent is false.

The model theory rests on an important principle concerning "working memory", which is the memory that holds a small amount of information, such as a telephone number, for a few seconds (unless one keeps rehearsing it). Such a memory for intermediate results is the heart of computational power. For example, more computational power is required to carry out multiplication than to carry out addition. Because working memory in human beings has a limited capacity, the model theory rests on the following principle of truth:

Individuals tend to minimize the load on working memory by constructing mental models that represent what is true, but not what is false.

The principle seems straightforward, but it has some subtleties illustrated by another exclusive disjunction:

There is a king in the hand or else there is not an ace in the hand.

It has two mental models:

King

¬ Ace

where "¬" denotes negation. The first model represents that there is a king, but it does not represent explicitly that it is false that there is not an ace in this situation. The second model similarly does not represent explicitly that it is false that there is a

king in this situation. People list the possibilities shown in the models above when they enumerate the different ways in which the assertion could be true. Falsity should not be confused with negation: falsity is a semantic notion, whereas negation is a syntactic one. Hence, a negative assertion can be true or false. Mental models do not normally represent falsity (unlike a truth table). And, models of the true possibilities represent only those literal propositions in the premises, such as "there is a king" and "there is not an ace" in the assertion above, that are true. The theory assumes that reasoners make "mental footnotes" about the propositions that are false in a situation, such as the falsity of "there is not an ace" in the first model above, but these footnotes are hard to remember. They are ephemeral, and reasoners soon lose track of them.

It is the principle of truth that accounts for the difficulty of the Chernobyl inference. Given premises of the form:

If A then B

Not B

reasoners construct models of the conditional premise:

A B

. . .

The second premise, not B, is not consistent with the first of these models, which it therefore eliminates. Only the implicit model is left, and because it has no content it seems that nothing follows. In order to draw the correct conclusion, it is necessary to flesh out the implicit model to represent explicitly what happens when A is false:

A B

¬ A B

¬ A ¬ B

where true negations are used to represent false affirmatives. The premise, not B, eliminates the first two models, and so the conclusion, not A, follows from the remaining model.

Formal rule theories have had some success in accounting for the difficulty of inferences in terms of the length of their formal derivations. The mental model theory has also had some success in accounting for the difficulty of inferences in terms of the numbers of explicit models that have to be constructed in order to carry them out. Unlike formal rule theories, however, the model theory also accounts for systematic invalid conclusions that naive individuals draw: they tend to correspond to just a single model of those premises that, in fact, have other models. Reasoners often fail to consider all the models of such premises, or fail to discern what is common to their multiple models. The model theory also has a wider purview because it provides a unified account of reasoning about what is possible, probable, and necessary. A conclusion is possible if it holds in at least one model of the premises; it is probable if it holds in most models of the premises; and it is necessary if it holds in all the models of the premises. These advantages of the model theory have not been universally conceded by its critics, and formal rule theories still have many active adherents. What is needed to decide between the two accounts is a crucial prediction -- a prediction made by one theory that contravenes the other theory -- and it is to such a prediction that we now turn.

A crucial prediction of "illusory" inferences

Readers should consider the following two problems, and write down their answers to each of them:

Problem 1: 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?

Problem 2: 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?

For problem 1, the model theory postulates that individuals consider the true possibilities for each of the three premises. For the first premise, they consider three models, shown here on separate lines, 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, reasoners should respond, "yes, it is possible for an ace to be in the hand". The second premise also supports the same conclusion. In fact, reasoners are failing to take into account that when, say, the first premise is true, the second premise:

There is a queen in the hand or there is an ace, or both

is false, and so there cannot be an ace in the hand. The conclusion is therefore a fallacy. Indeed, if there were an ace in the hand, then two of the premises would be true, contrary to the rubric that only one of them is true. The same strategy, however, will yield a correct response to a control problem in which only one premise refers to an ace. Problem 1 is an illusion of possibility: reasoners infer wrongly that a card is possible. A similar problem to which reasoners should respond "no" and thereby commit an illusion of impossibility can be created by replacing the two occurrences of "there is an ace" in problem 1 above with, "there is not an ace". A recent experiment examined the two sorts of illusion and their respective control problems (Johnson-Laird and Goldvarg, 1997), with half of the illusions based on disjunctive premises and half based on conditionals. The participants' confidence in their conclusions did not differ reliably from one sort of problem to another. Yet, they were highly susceptible to the illusions, and performed well with the control problems. For example, 99% of responses to problem 1 were the predicted illusory, "yes". 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 were less likely to succumb to the illusions of impossibility.

If the illusions result from a failure to represent what is false, then any manipulation that emphasizes falsity should reduce them. When individuals had to generate false instance of the premises prior to the main task, they were less susceptible to the illusions (Legrenzi and Girotto, 1996). Likewise, the rubric, "Only one of the following two premises is false," reliably reduced their occurrence. Unfortunately, people do not have a direct access to the cases in which disjunctions and other comparable assertions are false. They have to infer them from the situations in which they would be true.

With hindsight, it is surprising that nearly everyone responded "yes" to problem 1, because it seems obvious that an ace renders two of the premises true. A recent experiment tested two groups of participants, and half way through the experiment, one group was told to check whether their conclusions met the constraint that only one of the premises was true (Johnson-Laird and Goldvarg, 1997). This procedure has the advantage that the participants do not have to envisage the circumstances in which the premises would be false. The group that received the special instruction was indeed much less likely to commit the fallacies thereafter.

The rubric, "one of these assertions is true and one of them is false", is equivalent to an exclusive disjunction between two assertions:

A or else B, but not both

This usage leads to still more compelling illusions that seduce novices and experts alike. Consider problem 2 above. Nearly everyone infers that there is a king in the hand, which is the conclusion predicted by the mental models of the premises. This problem tricked the first author in the output of his computer program implementing the model theory. He thought at first that there was a bug in the program when his conclusion -- that there is a king -- failed to tally with the one supported by the fully explicit models. The program was right. The conclusion 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. If, say, the first conditional is false, then there need not be a king in the hand even though there is a jack. And so the inference that there is a king is invalid: the conclusion could be false.

An unpublished experiment carried out by Fabien Savary examined problem 2 and another illusion, and compared them with two control problems in which the neglect of false cases should not impair performance. 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 highly confident in both their illusory conclusions and their correct control conclusions, with no reliable difference between them.

Because so many experts have made 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 highly 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. The other putative explanations concern the meaning of conditionals, and their meaning is controversial. However, the illusions occur with disjunctions too, and their meaning is transparent.

The existence of systematic fallacies in human reasoning came as a shock to us. These illusions were predicted by the model theory, but they count against current theories of reasoning based on formal rules of inference. These theories rely solely on valid rules of inference and so cannot at present account for them. Many other robust phenomena in deductive reasoning appear to arise from the neglect of what is false. We will not pursue these cases any further, but rather turn to some phenomena in probabilistic reasoning, decision making, and recursive thinking, which may be relevant to other disciplines.

Reasoning about probabilities

Mental models suggest how people infer the probability of an event from their knowledge of the different ways in which the event can occur. The theory postulates that naive individuals assume by default that each model is equiprobable. They then infer the probability of an event from the proportion of models in which the event occurs. Consider this problem:

There is a box in which there is at least a red marble, or else a green marble and a blue marble, but not all three marbles. What is the probability that there is both a red and a blue marble in the box?

The premise elicits the following mental models:

Red

Green Blue

Participants in recent studies of such inferences tended to infer, as these models predict, a probability of zero for red and blue (Johnson-Laird, Legrenzi, Girotto, Legrenzi, and Caverni, 1998). The fully explicit models of the premises establish that where there is a red marble, there are three distinct ways in which it can be false that there is both a green marble and a blue marble:

Red Green ¬Blue

Red ¬Green Blue

Red ¬Green ¬Blue

¬Red Green Blue

Granted equiprobability, this partition shows that the probability of red and blue is 1/4. Naive reasoners succumb to a variety of other illusory inferences about probabilities (Johnson-Laird and Savary, 1996).

The model theory also accounts for inferences of posterior probabilities (Johnson-Laird et al, 1998). Consider the following example:

According to a population screening, a person has 4 out of 10 chances of having a certain disease; 3 out of the 4 chances of having the disease are associated with a particular symptom; 2 out of the 6 chances of not having the disease are also associated with the symptom. Mary is tested. Out of 10 chances, she has ___ chances of having the symptom, and among these chances, ___ chances will be associated with the disease.

The orthodox solution to the problem depends on using Bayes's theorem. However, naive individuals can build the following models of the premises:

disease symptom 3

disease ¬ symptom 1

¬ disease symptom 2

. . . 4

where each explicit model represents a separate possibility and the chances of its occurrence (out of 10), and the implicit model represents the 4 chances of having neither the disease nor the symptom. The posterior probability can be computed without recourse to Bayes's theorem, but by using a simple subset procedure. A conditional probability, p(A|B), depends on the subset of B that is A. In the model above, there are a total of 5 chances of having the symptom, and within them there are 3 chances of having the disease. Hence, the chances that Mary has the disease given that she has the symptom are 3/5. Girotto and Gonzalez (1998) have shown that naive reasoners tend to respond correctly to problems formulated so that they can construct such models and infer separately the denominator and numerator of the conditional probability.

Focusing in decision making

The classical theory of decision making, whatever its status as a specification of rationality, does not begin to explain the mental processes underlying decisions. On the one hand, the theory is radically incomplete: it has nothing to say about when one should decide to make a decision, or how one should determine the range of options and assess the utilities of their various outcomes. On the other hand, the theory conflicts with the evidence on how people reach decisions in daily life: their conspicuous failure to maximize expected utility has led some theorists to worry about human rationality and other theorists, notably Simon (1959), to argue for a different criterion for human decisions.

When individuals construct mental models, as we have seen, they make explicit as little as possible, and they focus on that information which is explicit in their models. Concomitantly, they fail to consider possibilities that lie outside their models. The consequence is that they may overlook the correct possibility. Many of the cognitive errors that have contributed to real-life disasters have exactly this form. For example, the operators at Three Mile Island explained the high temperature of a relief valve in terms of a leak, and overlooked the possibility that it was stuck open; the master of the English Channel ferry, The Herald of Free Enterprise, inferred that the bow doors had been closed, and overlooked the possibility that they had been left open; the engineers at Chernobyl found an erroneous explanation for the initial explosion and overlooked the possibility that the reactor had been destroyed.

In making decisions, focusing implies that individuals will often fail to make a thorough search for alternatives. In particular, if they are faced with the choice of whether or not to carry out a certain action, then they will construct a model of the action and an alternative model in which it does not occur. The latter will either be implicit or else merely a model in which the action does not occur. Hence, they will be focused on the action and search for more information about it in order to reach their decision. They will neglect to search for information about alternative actions, and so contravene the rational standard for decision making. Empirical evidence exists that people fail to consider alternative options and their costs when they are unstated (i.e. people seem to ignore the opportunity-cost principle, Friedman and Neumann, 1980). However, we can predict that focusing should be reduced by any manipulation that makes alternatives to the action more available.

We tested this prediction in an experiment in which we used a new approach to understanding the mental processes underlying decision making (Legrenzi, Girotto, and Johnson-Laird, 1993). The participants' task was to make a simple riskless decision, e.g. whether or not to go to see a certain movie. They could request any information that they needed in order to make the decision, and the experimenter gave it to them from one of two pre-established scenarios. The participants continued to request information until they were able to announce their decision. In the control group, the decision was presented without any background context, and so the participants were highly focused, that is, they requested information only about the action and ignored its alternatives. In the "context" group, the decision occurred within a particular context: the subjects were asked to imagine that they were visiting, say, Rome for the first time, and that the experimenter was an expert on the city's tourist attractions. The context made alternatives more available, and so the participants were less focused on the action in their requests for information to help them make the decision. The participants made five different decisions (in a random order), and there was a highly significant focusing effect, which occurred for all the participants in both groups. But, there was also a reliable difference between the two groups: none of the participants in the control group ever asked any questions about alternatives to the focused action, whereas 88% of the participants in the context group asked at least one question about alternatives to it.

We observe the same phenomenon in daily life. There is a natural tendency to focus when a single option is offered for a decision and no obvious alternatives are available. This result is contrary to any theory which assumes that decision makers explicitly consider alternative courses of action. In particular, it is contrary to the classical theory of decision making to which many economists still adhere. Focusing is important to our understanding of how the mind departs from rational principles: if one knows nothing about the alternatives to a particular course of action, one can neither assess their utilities nor compare them with the utility of the action. Hence, one cannot make a rational decision.

One unexpected finding was a close association between asking about alternatives to the focused action and deciding not to take the action. One reason may be that individuals who ask for and receive information about alternatives are thereby led to a negative decision. Another reason may be that those who are in the process of rejecting a course of action are thereby led to ask about alternatives. There may even by some other underlying factors that bias subjects to ask about alternatives and to reach negative decisions. These potential explanations are not mutually exclusive, but we suspect that the first of them is likely to be the main factor, because people reach negative decisions without asking about alternatives, whereas all but one instance of asking about alternatives led to a negative decision.

Focusing should lead to predictable requests for certain sorts of information in making decisions between two explicit alternatives. For example, suppose individuals have to choose between two alternative vacation resorts:

Resort A has good beaches, plenty of sunshine, and is easy to get to.

Resort B has good beaches, cheap food, and comfortable hotels.

What further information are individuals likely to request in order to choose between the two resorts? The focusing hypothesis implies that they will seek to build models that flesh out the missing values of the stated attributes. They know, for example, that resort A has plenty of sunshine, but they know nothing about the weather at resort B, and so they will seek information about this attribute. We can also predict that once the specified attributes have been fleshed out in this way, people will be able to make a decision provided that one resort dominates the other, i.e. it has all the positive attributes of the other, and some additional positive attributes. In summary, the initial specification of the decision acts as a focus for both the information that individuals seek and their ultimate decision, and consequently they will tend to overlook other attributes, such as hostility to tourists, a dictatorial government, or rampant food poisoning, that are not included in the original specification. Yet, these attributes might well influence their decisions in other instances. We note that one factor in the early success of Ross Perot in his bid for the U.S. Presidency in 1992 appears to have been the attractiveness of those attributes that he revealed in his television appearances. Those who supported him appeared to have focused on these attributes, and to have neglected others -- such as most matters of policy -- that he chose not to reveal.

The model theory’s predictions about reasoning can be carried over to decision making. Thus, a robust phenomenon in reasoning is that deductions that call for more than one model are harder than those that call for only one model. We have observed this phenomenon in studies of propositional, relational, syllogistic, and multiply-quantified reasoning (for a review, see Johnson-Laird and Byrne, 1991). Deductive performance can be tested to the point where it breaks down merely by increasing the number of disjunctive models that have to be constructed. We can expect that decisions will likewise grow more difficult as a function of the number of options. Indeed, like deduction, there can be a breakdown in rationality as soon as there are two explicit alternatives to choose from -- as shown by the so-called "disjunctive effect" in decision making (see Shafir and Tversky, 1992). An everyday example of this phenomenon occurred during the early stages of the 1992 US Presidential campaign: at one point, the opinion polls revealed that Bush would lose to a Democrat, but that he would beat

each individual Democratic candidate. In general, as Shafir and Tversky (1992) have shown experimentally, individuals may choose a particular option when a certain event occurs, and also when it does not occur. Yet, when the outcome is unknown, they do not choose it. In our terms, the need to hold in mind the disjunctive alternatives makes the inferential task difficult. A decision based on two alternatives, as Legrenzi and Girotto (1996) write, requires individual to build two explicit models and to assess what both of them have in common.

We can predict at least one other disjunctive effect in decision making: if the information available about a particular option is disjunctive in form, then the resulting conflict or load on working memory will make it harder to infer a reason for choosing this option in comparison to an option for which categorical information is available. The harder it is to infer a reason for a choice, the less attractive that choice is likely to be.

Models and meta-reasoning

Meta-reasoning is reasoning about other people's reasoning. As social psychologists, sociologists, and economists, have long understood, it plays a key part in human interactions. In public places, for instance, much of our deliberate behavior is designed to ensure that other people will infer that our activities are legitimate. A man loitering on a street corner will ostentatiously keep looking at his watch, and frowning in irritation, to make clear to the world at large that he is waiting for someone. He does not need to keep checking the time, but his action legitimizes his presence there. He has reasoned about what others are likely to reason and adjusted his behavior so that they will think what he wants them to think.

Meta-reasoning, we believe, depends on ordinary deductive powers, which it calls upon much as a computer program calls upon a subroutine to carry out some humdrum operation. It is also a source of logical puzzles, which have formed the main basis for its investigation in the psychological laboratory. Meta-logical puzzles depend on reference to truth and falsity, either directly or indirectly, whereas meta-reasoning puzzles depend on reasoning about how others have reasoned.

Here, for example, is a species of meta-reasoning puzzle that Johnson-Laird and Byrne (1991) devised:

There are two sorts of people: logicians, who always make valid deductions; and politicians, who never make valid deductions.

A says that either B is telling the truth or else B is a politician (but not both).

B says that A is lying.

C deduces that B is a politician.

Is C a logician?

The puzzle can be solved in the following way. Suppose that A is telling the truth, then there are two alternatives: B is telling the truth, or else B is a politician. But, B asserts that A is lying, and so the first alternative leads to a contradiction. Hence, if A is telling the truth, it follows validly that B is a politician. Now, suppose that A is not telling the truth. It is then not the case that either B is telling the truth or is a politician, i.e. it follows that B is telling the truth if and only if B is a politician. But, if B is not telling the truth, then A is not lying, i.e. A is telling the truth -- a consequence that contradicts the assumption. Hence, B must be telling the truth, and so it follows that B is a politician. Thus, whether A tells the truth or not, it follows that B is a politician. Because C deduces this valid conclusion, C must be a logician.

Problems that depend for their solution on deducing what one person can deduce about another person's deductions have been investigated in studies carried out by George Erdos of the University of Newcastle-on-Tyne. He gave his subjects meta-reasoning puzzles of the following well-known variety:

> Three wise men who were perfect logicians were arrested by the Emperor on suspicion of subversion. He put them to the following test. The three men were lined up facing in the same direction, and a hat was placed on the head of each of them. The men could not see, their own hats, but the man at the back of the queue (A) could see the two hats in front of him, the man in the middle (B) could see the one hat in front of him, and the man at the front (C) could see no hat. The Emperor said: "If one of you can tell me the color of your own hat, I will set all three of you free. There are three white hats and two black ones from which your hats have been drawn. I will now ask each of you if he can tell me the color of his hat. You may answer only ‘yes’, or ‘I don't know’." A who could see the two hats in front of him said, "I don't know". B heard A's answer and said, "I

don't know". C heard the two previous answers. What was C's answer?

The problem can be solved by considering the deductions made by the logicians. B deduces that if A had seen two black hats in front of him:

A B C

black black

A would have said "yes" as A would have known that his own hat must be white, because there were only two black hats in the set from which the Emperor made his selection. But since A said "I don't know", B concludes that A did not see two black hats, i.e. A must have seen one of the following three possibilities:

A B C

white white

black white

white black

C, in turn, deduces that if B had seen a black hat, then B would have said "yes" because, by the previous deduction, B would have known that the last of the three possibilities above was correct, and so his own hat must be white. Hence, B must have seen a white hat (and of course not known whether his own hat was black or white). Hence, C concludes that his own hat must be white, and he answers "yes" to the Emperor.

This type of recursive argument generalizes to any number, n, of individuals provided that the Emperor makes his selection from n white hats and n - 1 black hats. Thus, where there are four men wearing hats, D, who sees no hats can argue as follows: "If I am wearing a black hat, then the other three will see it and know that at most there are only two remaining black hats for them. This case is therefore identical to the three-hat problem. If none of the three has said "yes", then that implies that I have a white hat."

These meta-reasoning problems are hard to solve. The source of their difficulty is likely to be three-fold. First, the problems place a considerable load on working memory because a reasoner has to construct a model of one person's model of another person's model of the situation, which may depend in turn on yet another person's model. Thus, if you try to solve the three hats problem, you must first represent A's mental models, use them to infer B's mental models, and in turn use them to infer C's mental model. Such recursive considerations call for keeping track of the different individuals' models, and that in itself produces quite a load on working memory. Erdos reports that subjects often correctly deduce that A cannot see two black hats, but when they come to consider B's situation they forget that B can make this deduction too. Second, as in the problem above, it may be necessary to construct and to retain a disjunctive set of models, e.g. C's reasoning hinges on B's representation of the three possibilities that A could have seen. Disjunctive models are a source of difficulty because of their load on working memory. Third, the particular strategy to be adopted is not one that ordinary individuals are likely to be equipped with prior to the experiment. They have to reflect upon the problem and to discover for themselves the information latent in each of the logicians' answers. This task can be made easier, as Erdos has shown, by first giving the subjects a simple two-hat problem. Finally, it is worth noting that his participants did not find it helpful to be presented with a table of all the possible combinations of hats. In our view, this table swamps them with information.

Conclusions

The theory of mental models rests on simple principles, but, as we have tried to show, it leads to unexpected consequences. It extends in a natural way to inferring probabilities, to decision making, and to recursive reasoning about other people's reasoning. We can summarize the theory in terms of its three principal predictions, which have all been corroborated experimentally:-

1. Reasoners normally build models of what is true, not what is false -- a propensity that led to the discovery that people commit systematic fallacies in reasoning.
2. Reasoning is easier from one model than from multiple models.

3. Reasoners tend to focus on one of the possible models of multi-model problems, and are thereby led to erroneous conclusions and irrational decisions.

We should also add that the theory accounts for the informality of arguments in science and daily life, whereas logic is notoriously of little help in analyzing them. If people base such arguments on mental models, then there is no reason to suppose that they will lay them out like the steps of a formal proof. The theory of mental models, however, is not a paragon. It is radically incomplete; and it is likely to have problems and deficiencies. Proponents of rule theories have accused it of every conceivable short-coming from blatant falsehood to untestability. It postulates that human reasoners can in principle see the force of counterexamples, and indeed people are able to construct them (Bucciarelli and Johnson-Laird, 1998) -- a competence that is beyond the power of formal rule theories to explain. The model theory may well be overturned by counterexamples predicted by a superior theory. In which case, it will at least have had the virtue of accounting for its own demise.

To build models of what is true is a sensible way to deal 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. Without them, however, reasoners remain open to the illusion that they grasp what is in fact beyond them. We suspect that similar short-comings may underlie judgment and choice in game-theoretic settings.

References

Bucciarelli, M., and Johnson-Laird, P.N. (1998) Strategies in syllogistic reasoning. Under submission.

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

Friedman, L.A., & Neumann, B.R. (1980). The effects of opportunity costs on project investment decisions: A replication and extension. Journal of Accounting Research, 18, 407-419.

Girotto, V., and Gonzalez, M. (1998) Strategies and models in statistical reasoning. In Schaeken, W., and Schroyens, W. (Eds.) Strategies in Deductive Reasoning. In press.

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

Johnson-Laird, P.N., and Goldvarg, Y. (1997) How to make the impossible seem possible. Proceedings of the Nineteenth Annual Conference of the Cognitive Science Society, 354- 357.

Johnson-Laird, P.N., Legrenzi, P., Girotto, V., Legrenzi, M., and Caverni, J-P. (1998) Naive probability: a mental model theory of extensional reasoning. Unpublished MS, Princeton University.

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

Legrenzi, P., and Girotto, V. (1996) Mental models in reasoning and decision making processes. In Oakhill, J., and Garnham, A. (Eds.) Mental Models in Cognitive Science. Hove, Sussex: (Erlbaum UK) Taylor and Francis, Psychology Press, pp. 95-118.

Legrenzi, P., Girotto, V., and Johnson-Laird, P.N. (1993) Focussing in reasoning and decision making. Cognition, 49, 37-66.

Medvedev, Z. A. (1990) The Legacy of Chernobyl. New York: W.W. Norton.

Metzler, J., and Shepard, R.N. (1982) Transformational studies of the internal representations of three-dimensional objects. In Shepard, R.N., and Cooper, L.A. Mental Images and Their Transformations. Cambridge, MA: MIT Press. pp. 25-71.

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

Shafir, E., & Tversky, A. (1992). Thinking through uncertainty: nonconsequential reasoning and choice, Cognitive

Psychology, 24, 449-474.

Simon, H.A. (1959) Theories of decision making in economics and behavioral science. American Economic Review, 49, 253-283.

Wittgenstein, L. (1922) Tractatus Logico-Philosophicus. London: Routledge & Kegan Paul.

Return to the ICOS Home Page