

Human reasoning and rationality

by

Philip N. Johnson-Laird

November 27th 2001

A paper to be presented to the International Symposium on Foundations and the

Ontological Quest: Prospects for the New Millennium

Cognitive Sciences Section (Organizer: Sarah J. Nelson)

at

The Pontifical Lateran University, The Vatican, January 9th 2001

Author's address

Department of Psychology

Princeton University

NJ 08540, USA

Tel: (609) 258 4432

Fax: (609) 258 1113

Email: phil@princeton.edu

Introduction

Once upon a time I was sitting on a London tube train in Baker Street station. It was bound for Uxbridge. Just before the doors closed, a man came running onto the train and asked: Does this train go to Ickenham? I looked up at the map on the side of the train, and I found Ickenham a couple of stops before Uxbridge. And so I replied: "Yes". At that point, the doors closed. It occurred to me that the train might rush through the station without stopping, an event that does occur on certain parts of the tube system. But, I comforted myself with the thought that the man had asked whether the train went to Ickenham, not whether it stopped there.

Fortunately, the train did stop at Ickenham.

This episode illustrates a modest but successful case of reasoning. I knew the train was going to Uxbridge, and I saw from the map that any train that went to Uxbridge also went to Ickenham. As so my starting point was two premises:

The train goes to Uxbridge.

If a train goes to Uxbridge then it goes to Ickenham.

From which it follows:

The train goes to Ickenham.

The inference is valid, that is, if its premises are true, then its conclusion is must be true too.

Not all human reasoning is successful. Some years ago, the engineers in charge of an experiment knew two things:

If the experiment was to continue the turbines had to be rotating fast enough to generate emergency electricity.

The turbines were not rotating fast enough to generate emergency electricity.

The chances are that you, the reader, would infer:

The experiment was not to continue.

Unfortunately, the engineers failed to draw this conclusion. They continued the experiment, and it led to the Chernobyl disaster (see Medvedev, 1990). It is possible that the engineers did draw the conclusion, but for some reason failed to act on it. Yet, as experiments in the psychological laboratory show, the sort of inference about the tube train is easier than this second sort of inference – to which people often respond, "nothing follows" (see e.g. Evans, Newstead, and Byrne, 1993, for a review of these experiments).

A long-standing tradition in Western thought postulates that human thinking is infallible because it is founded on the "laws of thought". That is to say, formal rules of inference akin to those of logic guide human deductive reasoning, the probability calculus guides human reasoning about probabilities, and the principles of rational decision making guide human decisions. Unless we have been taught these principles, none of us is aware of any of them. They are therefore supposed to guide us unconsciously. Unfortunately, the evidence from daily life and from the psychological laboratory raises doubts about the hypothesis. The mere fact that humans make mistakes is not enough to refute the hypothesis, because mistakes could be haphazard and attributable to

momentary errors in performance – spanners in the works rather than flawed principles. But, this chapter will show, that claim is wrong. The bad news is that individuals make systematic and predictable errors in reasoning, and sometimes these errors are so compelling that nearly everyone makes one and the same error. The good news is that these errors are explicable, and they reveal the inner workings of the mind.

That people err in reasoning might be considered of no more importance than, say, the fact that they err in predicting the future. Deductive reasoning, however, is a central component of human intelligence. Without it, there would be no mathematics or science, no legal or ethical systems, and no cultural or social norms. (This article would also not exist.) Moreover, the conundrum of how people reason is an ideal "test" case for cognitive science. It is central to thinking, but almost certainly easier to understand than creativity. And a proper understanding of the mental capacity for reasoning is likely to lead to the development of methods to improve the ability, and thereby perhaps to reduce the risk of the sorts of errors that caused the disaster at Chernobyl.

The aims of this article are threefold. First, it outlines a theory of human reasoning that is a viable alternative to the "laws of thought". Second, it describes some characteristic evidence that supports this theory. Third, it draws some general morals about the nature of human rationality.

The theory of mental models

Sherlock Holmes, the great detective in the stories and novels of Sir Arthur Conan Doyle, once described his method of thought in the following words:

When you have eliminated the impossible, then whatever remains (however improbable) must be the case.

This idea makes excellent sense. It anticipates a theory of reasoning that my colleagues and I have advanced and gradually refined over the years (see, e.g., Johnson-Laird, 1983; Johnson-Laird and Byrne, 1991, 2001). This theory postulates that reasoning depends on understanding the meaning of premises, and then using this meaning and general knowledge to construct mental models of the possibilities under description. Three main assumptions distinguish the theory from other theories and, in particular, from latter day versions of the "laws of thought" (Rips, 1994; and Braine and O'Brien, 1998).

Possibilities lie at the heart of the model theory, and the first assumption relates them to models:

1. A mental model represents a possibility.

A model captures what is common to the different ways in which the possibility might occur. Like a diagram, the model is iconic, i.e., its parts correspond to the parts of what it represents, and its structure corresponds to the structure of the possibility. As an example, consider an assertion (in the form known as a "disjunction") about a fault in a nuclear power station:

The fault is in the valve or in the turbine, or both.

It calls for three mental models to represent three possibilities, which can be laid out on separate rows:

valve	
	turbine
valve	turbine

where "valve" denotes a mental model of the possibility in which the fault is in the valve, "turbine" denotes a model of the possibility in which the fault is in the turbine, and the final possibility allows that the fault is in both. Individuals list these three possibilities when they are asked to state what is possible given the assertion.

It is necessary to use words, such as "valve" and "turbine", to illustrate models. But, the reader should not confuse the diagram above with what it denotes, i.e., a set of mental models. In general, they can represent three-dimensional entities, spatial relations, temporal relations, events, processes, complex systems, and even abstract ideas. They can represent discourse about real, hypothetical, or imaginary cases; and they can reside in long-term memory as a representation of knowledge (see, eg., Johnson-Laird and Byrne, 2001). They underlie visual images, though many components of models are not visualizable.

The models of the assertion above illustrate the second principle of the theory, which is so important that it is dignified with a name:

2. The principle of truth: mental models represent what is true, but by default not what is false.

As the example shows, mental models represent only the possibilities that are true given an assertion. At a lower level, however, a model represents a clause in a description, or set of premises, only when the clause is true in the possibility. For example, the first model of the assertion above represents that the flaw is in the valve, but it does not represent explicitly that in this possibility it is false that the flaw is in the turbine. The principle of truth postulates that individuals by default do not represent what is false. But, there are exceptions. Individuals make "mental footnotes" about the falsity of clauses, and if they retain the footnotes they can flesh out mental models into fully explicit models, which represent clauses even when they are false. The following fully explicit models, for example, represent the earlier assertion:

Valve	\neg Turbine
\neg Valve	Turbine
Valve	Turbine

where the symbol " \neg " represents negation. The principle of truth postulates that people normally do not represent what is false; and it is important not to confuse falsity with negation. People will represent a negative assertion provided that it is true. Although they can represent fully explicit models, the principle of truth is the norm. It makes for parsimonious representations, because reasoners do not have to bother with what is false, and in this way they reduce what they have to hold in mind as they think. Psychologists refer to the system that holds information in mind as "working memory", and so the principle of truth reduces the load on working memory.

So far, we have seen how individuals can represent information presented to them in descriptions: they imagine the corresponding possibilities in the world. They may also construct such representations on the basis of perception, memory, and imagination. But, how do they use mental models in order to reason?

The answer to this question derives from the concept of a valid inference. Readers will recall that an inference is valid only if its conclusion must be true given that its premises are true. In other words, there are no counterexamples to the conclusion: it holds in every possibility compatible with the premises. As an example, consider the following valid inference:

The fault is in the valve or in the turbine, or both.

The fault is not in the turbine.

Therefore, the fault is in the valve.

The first premise yields three mental models of the possible locations of the flaw:

valve

turbine

valve

turbine

The second premise eliminates the last two of these models, and so only the first model remains:

valve

This model yields the conclusion that the fault is in the valve; and this conclusion is valid because it holds in all the models – in this case, the single model – of the possibilities compatible with the premises.

In contrast, consider the following inference:

The fault is in the valve or in the turbine, or both.

The fault is in the valve.

Therefore, the fault is not in the turbine.

The second premise eliminates just one model, but in order to do so, the mental models above do not suffice. It is necessary to consider the fully explicit models of the first premise. These models can be constructed only if reasoners have held onto the mental footnote about what is false. They can then flesh out the mental models into fully explicit models of the three possibilities:

valve	\neg turbine
\neg valve	turbine
valve	turbine

It is difficult to hold in mind these three possibilities, but at the very least reasoners need to represent the second model in a fully explicit way so that they appreciate that the second premise eliminates it. There remain just the first and the third of the fully explicit models:

valve	\neg turbine
valve	turbine

The third model, however, is a counterexample to the putative conclusion. It shows that the premises can be true, but the conclusion false. The fault can be in the turbine and the valve. Hence, the inference is invalid.

The model theory naturally extends to conclusions, not about matters of fact, but about possibilities. In the preceding example, the first of the two models of the premises support the following conclusion:

It is possible that the fault is not in the turbine.

Conclusions about a possibility are valid provided that they hold in at least one model (of a possibility) compatible with the premises. Hence, this conclusion is valid.

Some Italian colleagues and I have extended the theory to reasoning about probabilities, when the probability of an event is based on all the different ways in which that event can occur (Johnson-Laird, Legrenzi, Girotto, Legrenzi, and Caverni, 1999). To take a simple example, consider the following question:

Given that the fault is in the valve or in the turbine or both, what is the probability that it is in the turbine?

If individuals know nothing more about the situation, then they are likely to respond: $2/3$. Reasoners tend to infer that the probability of an event is equal to the proportion of equiprobable cases in which it occurs. The English economist John Maynard Keynes tried to integrate deductive reasoning and probabilistic reasoning within a single theory (Keynes, 1921). But his theory was flawed (Ramsey, 1931). The model theory may provide a workable account that integrates the psychological processes of deductive and probabilistic reasoning.

The principles that I have illustrated for reasoning about possible, probable, and necessary conclusions are summarized in the third assumption of the theory:

3. Human reasoning depends on mental models.

Models can be used for reasoning according to the rational principle that a conclusion is valid if it holds in all the models of the premises, i.e., it has no counterexamples, and so it is necessary given the premises (Johnson-Laird and Byrne, 1991). If a conclusion holds in a proportion of models, its probability is equal to that proportion granted that the models represent equiprobable alternatives (Johnson-Laird, et al., 1999). If a conclusion holds in at least one model, it is possible given the premises (Bell and Johnson-Laird, 1998). And if it holds in none of the models, it is impossible given the premises.

The theory assumes that people normally reason using mental models, but that with simple tasks and simple assertions they can flesh out their models to make them fully explicit. A suite of computer programs, which I have written, implements the theory for a variety of domains. The programs construct mental models and fully explicit models, and they show how inferences can be made using models. The crucial issue to which we now turn is whether logically-untrained individuals use mental models in order to reason.

The evidence for models

The theory of mental models makes several predictions about human reasoning. Many investigators have corroborated them for a variety of sorts of reasoning (for a recent review, see Johnson-Laird, 2001). This section of the article describes three principal discoveries, which suffice to modify our picture of human rationality.

One model is better than many

The first, and perhaps most obvious, prediction of the model theory is that inferences that call for just one mental model should be easier than those that call for multiple models. They should take less time and be less prone to error. Conversely, the more models of possibilities which individuals have to hold in mind, the harder it should be to reason. The prediction appears to be true. Here is an example for you to try:

Maria is in Milano or else Vittorio is in Vicenza, but not both.

Vittorio is in Vicenza or else Paolo is in Perugia, but not both.

What follows?

The problem is quite hard (see Bauer and Johnson-Laird, 1993). You can envisage the two possibilities compatible with the first assertion:

Maria in Milano

Vittorio in Vicenza

But, it is difficult to incorporate the possibilities from the second assertion. In fact, the solution is as follows:

Maria in Milano

Paolo in Perugia

Vittorio in Vicenza

Hence, there are only two models of possibilities, and an emergent conclusion is:

Either Maria is in Milano and Paolo is in Perugia or else Vittorio is in Vicenza.

Given a big enough increase in the number of models, the human inferential system collapses under the load. The following problem defeats many people:

Maria is in Milano or Vittorio is in Vicenza, or both.

Vittorio is in Vicenza or Paolo is in Perugia, or both.

What follows?

The first assertion is consistent with three possibilities:

Maria in Milano

Vittorio in Vicenza

Maria in Milano

Vittorio in Vicenza

The task of adding the three possibilities compatible with the second assertion is very hard. The correct answer is:

Maria in Milano

Paolo in Perugia

Vittorio in Vicenza

Vittorio in Vicenza

Paolo in Perugia

Maria in Milano

Vittorio in Vicenza

Maria in Milano

Vittorio in Vicenza

Paolo in Perugia

An emergent conclusion from these five possibilities is:

Maria is in Milano and Paolo is in Perugia, or Vittorio is in Vicenza.

A super-human reasoner – an angel, perhaps – would grasp at once that the two preceding assertions yield five possibilities. Such a reasoner is embodied in my computer programs so that they can check what the correct conclusions are. Yet, any finite organism including a computer will be defeated by an explosive growth in the number of possibilities. Reasoning which hinges on such sentential connectives as "if", "or", and "and", is computationally intractable. What this jargon means is that as the number of possibilities goes up, so

eventually no computer – not even one as big as the universe running at the speed of light – can cope. It is a striking fact that most of us human reasoners cannot cope with more than a handful of possibilities. The task can be made easier if, instead of verbal premises, diagrams make the possibilities explicit (Bauer and Johnson-Laird, 1993). Even diagrams, however, will soon overwhelm finite reasoners.

Knowledge and beliefs influence reasoning

The second prediction of the model theory is that the content of inferences and background knowledge should influence reasoning. They certainly influence the interpretation of premises. Consider, as an example, the following problem:

Maria is in Milano or at least she's in Italy.

Maria is not in Italy.

What follows?

Most individuals (in an unpublished experiment carried out in collaboration with Tom Ormerod) draw the following conclusion:

Maria is not in Milano.

Psychological theories based on the laws of thought postulate the following formal rule of inference:

A or B, or both.

Not-B.

Therefore, A.

where A and B can denote any propositions whatsoever. But, the conclusion above corresponds to: Not-A. It is contrary to any formal rule of inference. Yet it is a sensible conclusion. So, what is going on?

An assertion of the form, A or B or both, is normally compatible with three fully explicit possibilities. Hence, the first assertion in the inference above should be compatible with the following three fully explicit possibilities:

Maria-in-Milano	\neg Maria-in-Italy
\neg Maria-in-Milano	Maria-in-Italy
Maria-in-Milano	Maria-in-Italy

General knowledge, however, eliminates the first possibility: it is impossible for Maria to be in Milano but not in Italy, because Milano is in Italy. Hence, no-one (other than a psychotic or a logician) is likely to make the inference governed by the formal rule:

Maria is in Milano or at least she's in Italy.

Maria is not in Italy.

Therefore, she's in Milano.

In contrast, the following sort of valid inference is easy (see Bouquet, and Warglien, 1999):

Paolo is in Perugia or else in Venezia.

He's not in Venezia.

Therefore, he's in Perugia.

You know that one person cannot be in two places at the same time, and so it is easy to infer that if Paolo is not in one place, then he is in the other.

Knowledge and beliefs also influence the process of reasoning. Consider, for instance, the following problem about some Frenchmen in a restaurant:

All the Frenchmen in the restaurant are gourmets.

Some of the gourmets are wine-drinkers.

What follows?

When we gave this problem to some students at Sussex University (see Oakhill, Garnham, and Johnson-Laird, 1990), the majority of them (78%) drew the conclusion:

Some of the Frenchmen are wine-drinkers.

This conclusion is plausible. Indeed, a separate group of individuals from the same population had rated it as highly believable. Now consider a contrasting problem:

All the Frenchmen in the restaurant are gourmets.

Some of the gourmets are Italians.

What follows?

Despite the salience of the European Union, the participants in the experiment who received this problem tended not to draw the conclusion of the same form as before, namely:

Some of the Frenchmen are Italians.

Only 8% of them drew this conclusion. The others rejected it. They knew that it was implausible; and our independent group of participants rated it as highly unbelievable. The model theory accounts for the phenomena. The participants

given the first problem constructed a single mental model of some individuals who satisfied the two premises:

Frenchman	gourmet	wine-drinker
Frenchman	gourmet	wine-drinker
Frenchman	gourmet	

Each row represents a different individual, and so there are three Frenchmen who are all gourmets, but only two of them are wine-drinkers. This model supports the credible conclusion:

Some of the Frenchmen are wine-drinkers.

And so reasoners are happy to draw this conclusion. In contrast, the analogous model of the second problem yields the incredible conclusion:

Some of the Frenchmen are Italians.

Hence, the participants search harder for a counterexample. Because the inference is invalid, they are likely to discover one:

Frenchman	gourmet	
Frenchman	gourmet	
	gourmet	Italian
	gourmet	Italian

Both premises are true in this model: all the Frenchmen are gourmets, and some of the gourmets are Italians. But, now none of the Frenchmen is an Italian, and so the model is a counterexample to the conclusion.

In general, individuals search harder for counterexamples to preposterous conclusions. This search is compatible with a robust finding: knowledge and

beliefs have a bigger effect on invalid inferences than on valid inferences (e.g. Evans, Barston, and Pollard, 1983; Cherubini, Oakhill, and Garnham, 1999).

There is some controversy about whether individuals use counterexamples in their reasoning. Recent research has shown that people do not all use the same fixed procedure in reasoning. They develop various ways to tackle inferential problems, and the same person may use different strategies from one problem to another (see, e.g., Schaeken, De Vooght, Vandierendonck, and d'Ydewalle, 2000). Nevertheless, certain problems elicit counterexamples from nearly everyone, e.g.:

More than half of the people at this conference speak Italian.

More than half of the people at this conference speak English.

Does it follow that more than half of the people at this conference speak both Italian and English?

Reasoners spontaneously draw diagrams, such as the one in Figure 1, to refute the conclusion (Neth and Johnson-Laird, 1999). A recent study (Kroger, Cohen, and Johnson-Laird, 2001) using functional Magnetic Resonance Imaging allowed us to determine which regions of the brain are active during deductive reasoning. Reasoning elicits much more activity in the right frontal lobes of the brain than mental arithmetic based on the same premises. These regions are likely to mediate the representation of spatial and abstract relations, and so the model theory predicts such an increase in activity. A striking result is shown in Figure 2: only those reasoning problems likely to elicit a counterexample activated the region in the right frontal lobe known as the frontal pole. This region is active

during the processing of conflicts, and a counterexample yields a conflict: it is compatible with the premises, but incompatible with the conclusion.

Insert Figures 1 and 2 about here

Falsity and fallacies

Readers are invited to solve the following problem and to write down their answer for future reference:

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 10, or both.

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

The principle of truth predicts that individuals consider the true possibilities for each assertion. For the first assertion, they consider three mental models, which each correspond to a possibility given the truth of the assertion:

king

ace

king ace

Two of the models show that an ace is possible. Hence, individuals should respond, "yes". The reader, I suspect, may well have made the same response.

Yet, it is wrong. It is impossible for an ace to be in the hand, because both of the first two assertions would then be true, contrary to the rubric that only

one of them is true. The problem is an illusion of possibility: reasoners infer wrongly that a state of affairs 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 the preceding problem with, "there is not an ace". To verify that individuals are reasoning, a control problem can be constructed by pairing the assertions above with the question:

Is it possible that there is a jack in the hand?

Here, reasoners should respond: "Yes"; and the response is correct. In an experiment, students at Princeton succumbed to the illusions but did well with the control problems (Goldvarg and Johnson-Laird, 2000).

Readers should now consider the following problem:

Could both these assertions be true about the same tray?

The tray is heavy or else it is not both elegant and portable.

The tray is heavy and not elegant.

The problem calls for reasoning about whether or not a set of assertions makes up a consistent description. According to the model theory, people carry out this task by searching for a single model of a possibility in which each of the assertions is true. If they find such a model, the description is consistent; otherwise, it is inconsistent. Most people say: "Yes" to the problem above. But, the response is an illusion, which the principle of truth predicts. Mental models represent what is true, not what is false, and so those for the first assertion are as follows:

heavy

\neg elegant portable

elegant \neg portable

\neg elegant \neg portable

The second assertion describing the tray is consistent with the first of these models in which the tray is heavy but not elegant, granted that the absence of information in a model about a property is treated as its negation. Hence, reasoners should respond that the two assertions are consistent. But, the mental models fail to take into account that when one clause in the first assertion is true, the other clause is false. In contrast, the fully explicit models of the disjunction, which take falsity into account, are as follows:

heavy elegant portable

\neg heavy \neg elegant portable

\neg heavy elegant \neg portable

\neg heavy \neg elegant \neg portable

These models show that the correct response is that the two assertions are inconsistent. The problem should therefore yield an illusion of consistency. A control problem pairs the same disjunction with a different conjunction:

Not-elegant and notportable.

It corresponds to the fourth of the mental models above, and so reasoners should respond that the two assertions are consistent. As the fourth of the fully explicit models shows, this response is correct. The principle of truth also predicts the occurrence of illusions of inconsistency in which reasoners infer that

a description is inconsistent when, in fact, it is consistent, e.g., the same disjunction as above is paired with the assertion:

Heavy, elegant, and portable.

The mental model of this conjunction does not occur among those for the disjunction, and so reasoners should respond that the two assertions are inconsistent. Yet, as the fully explicit models of the disjunction show, the response is an illusion. Finally, a control for this illusion pairs the previous disjunction with the conjunction:

Not-heavy, elegant, and portable

which does not match either the mental models or the fully explicit models.

In an experiment with an unusually large number of participants, 489 Italian high school graduates carried out 16 different problems, which included both sorts of illusion and their control problems (Legrenzi, Girotto, and Johnson-Laird, 2001). The participants tended to get the control problems right, but the illusory problems wrong: 459 of them did so, 11 participants went against this trend, and the remaining 19 did equally well, or badly, with the two sorts of problem. To establish the reality of a phenomenon, psychologists calculate the probability of its occurrence by chance if the experimental manipulation had no real effect. In this case, the results would occur by chance less than 5 times in 10^{99} experiments – a singularly improbable event.

Monica Bucciarelli of the University of Torino has discovered an illusion that concerns what people ought to do, i.e., so-called "deontic" reasoning (see

Bucciarelli and Johnson-Laird, 2001). Consider, for instance, the following problem:

You are permitted to carry out only one of the following two actions:

To take the apple or the orange, or both.

To take the pear or the orange, or both.

Are you permitted to take the orange?

The rubric to this problem implies that you should carry out either the action described in the first sentence or else the action described in the second sentence, but not both actions. The mental models of the first assertion represent what it is permissible to take:

Apple

Orange

Apple

Orange

and so it seems permissible to take the orange. The mental models of the second disjunction support the same conclusion. Hence, the model theory predicts that reasoners should respond: "yes". The response is an illusion. If you take the orange then you will have carried out both actions, contrary to the rubric that you are permitted to carry out only one of them. But, suppose that we change the rubric so that it concerns a prohibition:

You are prohibited from carrying out more than one of the following actions:

To take the apple or the orange, or both.

To take the pear or the orange, or both.

Are you permitted to take the orange?

When individuals think about what is prohibited, what comes to mind first, as we showed independently, is what they must not do. Hence, reasoners should construct the mental models of the cases that are impermissible, i.e., the cases when both actions are carried out. These cases include taking the orange, and so reasoners should realize that the answer is: No, they are not permitted to take the orange. They can make this inference as soon as they construct a model of the impermissible case of taking the orange. It follows that the illusory inferences should be greater for permissions than for prohibitions. Indeed, the experiment corroborated this prediction: reasoners tended to succumb to the illusion when the rubric concerned permission, but to resist it when the rubric concerned prohibition.

Many cognitive scientists have themselves fallen for illusory inferences. They have then proposed ingenious explanations for their mistake. For example, the premises are so ambiguous, artificial, or unusually worded, that they confuse people. This hypothesis overlooks the fact that reasoners are highly confident in their conclusions, and that the control problems, which people get right, are equally ambiguous, artificial, and unusually worded. The principle of truth predicts the illusions: where falsity matters, fallacies occur.

Conclusions

Does it matter that we all tend to make mistakes in reasoning? The answer is: yes, it matters both in theory and in practice. It matters in practice, because human errors can be costly. The single biggest limitation in human performance is our inability to make correct inferences. When a complex system goes wrong, its human operators are often unable to infer either the cause or the cure. A typical failure occurs when individuals infer the wrong model of the situation, and stick with it despite all the evidence to the contrary. The operators at Chernobyl were convinced that the reactor was still intact after the explosion, and their difficulty in reaching the correct conclusion led to a disastrous delay in evacuating personnel (Medvedev, 1990). In an ambiguous and stressful situation, the danger is that we build the wrong model, interpret new information in the light of this model, and continue in this way until our private reality conflicts horribly with the true situation (Perrow, 1984).

Sometimes there are too many possible models for individuals to infer which is the correct one. In less than a minute after the turbine tripped at Three Mile Island on March 28th 1979, there were more than a hundred alarm signals, many instruments had gone off their scales, and the computer printer was lagging over an hour behind the messages waiting to be printed. It was not until two hours later that the operators were able to infer the correct model of what had happened — a pressure-operated relief valve was stuck open.

A more prosaic example of how the number of possibilities can overwhelm comes from a typical U.K. government leaflet. It purports to explain the Death grant (a grant to help to defray the cost of a spouse's burial):

Death grant is payable where either of the following conditions is satisfied by the person on whose [National Health] contributions the grant is claimed:

- The contributor must have paid or been credited with at least 25 contributions of any class at any time between 5 July 1948 or the date of entry into insurance, if later, and 5 April 1985, or the date on which he reached 65 (60 for a woman), or died under that age, whichever is the earliest; or
- Since 6 April 1985 the contributor must have actually paid contributions in any one tax year before the relevant year, on earnings of at least 25 times the lower earning limit for that year. The relevant year is usually the income tax year in which the death occurred, but if immediately before the date of death, the person on whose contributions the grant is claimed was himself dead or over 65 (60 for a woman), it is either the year in which he reached that age, or the year in which he died, whichever is earlier.

A former British Minister of Health, the late Richard Crossman, was asked: Why can't Government leaflets be simple to understand? He immediately rejected the idea: if people understood them, he explained, then they would get the money to which they were entitled, and that would cost the Government too much.

The moral of these practical examples is twofold. On the one hand, technology must be designed – Government leaflets, apart – so that people can understand it, both when it works and when it goes wrong. On the other hand,

efforts must be made to improve human reasoning. Our recent understanding of human reasoning can indeed be exploited pedagogically. My former student, Victoria Bell, showed that a simple procedure makes a substantial improvement in the reasoning of naïve individuals. She merely asked them to enumerate the possibilities compatible with the premises (Bell, 1999).

Mistakes in reasoning matter in theory, because they modify our picture of human rationality. Once, humans were supposed to be logically impeccable, because the laws of thought guide them. We now know that this view is mistaken. Humans are not impeccable: they do not reason like angels, but make systematic and predictable errors. They rely, not on the laws of thought, but on mental models. This theory is based on three assumptions: each mental model represents a possibility, it represents only what is true, and the strength of a conclusion depends on the proportion of models in which it holds. The theory makes a number of predictions, and this article has reviewed the evidence for three of them:

1. One model is better than many. That is, the fewer the models needed for an inference, and the simpler they are, the easier is the inference.
2. Content and background knowledge influence the process of reasoning. Individuals search more assiduously for counterexamples to unbelievable conclusions.
3. When falsity matters, fallacies occur. The resulting illusions are seductive, and they occur in a variety of domains. Certain procedures alleviate them, but no-one has discovered a perfect antidote to them.

The laws of thought are embodied in current theories based on formal rules of inference. These theories make no use of possibilities, and they cannot account for the three preceding phenomena. Yet formal rules and mental models are not incompatible in principle. As reasoners develop, they may learn to construct formal rules for themselves in certain idealized domains. Such a step was essential for the development of formal logic as a discipline.

If we are not among the angelic orders as reasoners, are we wholly irrational? Certainly, to err is human. Yet, our case is not hopeless. Our central thread of rationality hangs from a simple principle. We grasp that an inference is good if it has no counterexamples. The model theory is founded on this principle. The theory contains many lacunae and is far from complete. It is even conceivable that it will be overturned by the discovery of robust counterexamples to its own predictions. If so, it will at least account for its own demise.

Acknowledgements

Preparation of this article was supported by a grant from the National Science Foundation (Grant 0076287) to study strategies in reasoning. The article was made possible by the community of reasoning researchers, and particularly those in Italy: Bruno Bara (Torino), Monica Bucciarelli (Torino), Paolo Cherubini (Padova), Vittorio Girotto (Trieste and CNRS, Aix-en-Provence), Maria Sonino Legrenzi (Padova), Paolo Legrenzi (Venezia), Patrizia Tabossi (Trieste),

and Massimo Warglien (Venezia). Many other researchers, too numerous to name, have also helped, and the author thanks them all.

References

- Bauer, M.I., and Johnson-Laird, P.N. (1993) How diagrams can improve reasoning. Psychological Science, 4, 372-378.
- Bell, V. (1999) The model method: An aid to improve reasoning. Unpublished Ph.D. dissertation, Department of Psychology, Princeton University, Princeton, NJ 08544, USA.
- Bell, V., and Johnson-Laird, P.N. (1998) A model theory of modal reasoning. Cognitive Science, 22, 25-51.
- Bouquet, P. and Warglien, M. (1999). Mental models and local models semantics: the problem of information integration. Proceedings of the European Conference on Cognitive Science (ECCS'99), Siena: University of Siena. Pp.169-178.
- Braine, M.D.S., and O'Brien, D.P., Eds. (1998) Mental Logic, Mahwah, NJ: Erlbaum.
- Bucciarelli, M., and Johnson-Laird, P.N. (2001) Deontic meaning and reasoning. Under submission.
- Cherubini, P., Oakhill, J.P., and Garnham, A. (1999) Can any ostrich fly? Some new data on belief bias in syllogistic reasoning. Cognition, 69, 179-218.
- Evans, J. St.B.T., Barston, J.L., and Pollard, P. (1983) On the conflict

- between logic and belief in syllogistic reasoning. Memory & Cognition, 11, 295-306.
- Evans, J. St. B. T., Newstead, S.E., and Byrne, R. M. J. (1993) Human Reasoning: The Psychology of Deduction. Mahwah, NJ: Erlbaum.
- Goldvarg, Y. and Johnson-Laird, P.N. (2000) Illusions in modal reasoning. Memory & Cognition, 28, 282-294.
- Jeffrey, R. (1981) Formal Logic: Its Scope and Limits, 2nd Ed., McGraw-Hill.
- Johnson-Laird, P.N. (1983). Mental Models: Towards a Cognitive Science of Language, Inference and Consciousness. Cambridge, Cambridge University Press. (Italian translation: I modelli mentale, trans. A. Mazzocco., Bologna: Il Mulino.)
- Johnson-Laird, P.N. (2001) Mental models and deduction. Trends in Cognitive Science, 5, 434-442.
- Johnson-Laird, P.N., and Byrne, R. M.J. (1991) Deduction. Hillsdale, N.J.: Erlbaum.
- Johnson-Laird, P.N., and Byrne, R.M.J. (2001) Conditionals: a theory of meaning, pragmatics, and inference. Psychological Review, In press.
- Johnson-Laird, P.N., Legrenzi, P., Girotto, V., and Legrenzi, M. (2000) Illusions in reasoning about consistency. Science, 288, 531-532.
- Johnson-Laird, P.N., Legrenzi, P., Girotto, V., Legrenzi, M., and Caverni, J-P. (1999) Naive probability: a mental model theory of extensional reasoning. Psychological Review, 106, 62-88.

- Keynes, J.M. (1921) A Treatise on Probability. London: Macmillan.
- Kroger, J.K., Cohen, J.D., and Johnson-Laird, P.N. (2001) A double dissociation between logic and mathematics. Under submission.
- Legrenzi, P., Girotto, V., and Johnson-Laird, P.N. (2001) Models of consistency. Under submission.
- Medvedev, Z. A. (1990) The Legacy of Chernobyl. New York: W.W. Norton.
- Neth, H. and Johnson-Laird, P.N. (1999) The search for counterexamples in human reasoning. Proceedings of the Twenty First Annual Conference of the Cognitive Science Society, 806.
- Oakhill, J., Garnham, A., & Johnson-Laird, P. N. (1990). Belief bias effects in syllogistic reasoning. In K.J. Gilhooly, M.T.G., Keane, R.H. Logie, & G. Erdos (Eds.), Lines of Thinking, Vol. 1. New York: Wiley. 125-138.
- Perrow, C. (1984) Normal Accidents: Living with High-risk Technologies. New York: Basic Books.
- Ramsey, F.R. (1931) Foundations of Mathematics. London: Routledge & Kegan Paul.
- Rips, L.J. (1994) The Psychology of Proof. Cambridge, MA: MIT Press.
- Schaeken, W., De Vooght, G., Vandierendonck, A., and d'Ydewalle, G., (Eds.) Deductive Reasoning and Strategies. Mahwah, N.J.: Erlbaum.

Legends for Figures

Fig 1: A typical diagram of a counterexample. Each x represents an individual, more than half of them speak Italian, more than half of them speak English, but it is not the case that more than half speak both languages.

Fig 2: The region of the brain (the right frontal pole, shown as a reddish area in the "slice" through the brain on the left) activated by reasoning problems eliciting a search for counterexamples, but not by other simple logical problems or mental arithmetic (from Kroger, Cohen, and Johnson-Laird, 2001). The results are from a functional Magnetic Resonance Imaging study of 16 human participants. The graphs on the right show the time course of activation for four sorts of problem: logical problems calling for counterexamples (the pink curve), easy logical problems (the red curve), hard mental arithmetic problems (the light blue curve), and easy mental arithmetic problems (the dark blue curve). The logical and arithmetical problems were based on the same premises. The vertical gray bar demarcates an interval of 8 seconds. The peak of activation for the problems calling for counterexamples was several seconds after the participants began solving the problems, which is consistent with manipulation of mental models.

Fig. 1

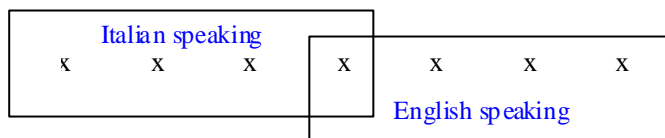


Fig. 2

