Perspectives

Human and computer reasoning

Philip N. Johnson-Laird

This paper reviews three main sorts of computer program designed to make deductive inferences: resolution theorem-provers and goal-directed inferential programs, implemented primarily as exercises in artificial intelligence; and natural deduction systems, which have also been used as psychological models. It argues that none of these methods resembles the way in which human beings usually reason. They appear instead to depend, not on formal rules of inference, but on using the meaning of the premises to construct a 'mental model' of the relevant situation, and on searching for alternative models of the premises that falsify putative conclusions. Experimental evidence corroborates this account of human reasoning.

The ability to reason is central to human life because without it there would be no science or culture, and no law or morality. Yet, despite its centrality, it is difficult to study. We are not aware of its underlying processes – we cannot stick microelectrodes into them, and even a detailed neural wiring diagram would reveal no more about them than the architecture of a computer reveals about the program that it is running. We cannot observe the inferential processes themselves, but only their consequences in speech and behaviour. Nevertheless, in recent years computer scientists have developed an increasing number of programs that reason, and psychologists have proposed theories of reasoning that for the first time are sufficiently explicit to be modelled in computer programs. This article reviews these two related fields of work.

It is, of course, futile to search for a comprehensive definition of a notion such as reasoning that is to be elucidated by a theory, but it is useful to demarcate the area of study with a working definition. In general terms, an *inference* refers to some systematic process of reasoning by which one set of propositions (the 'premises') leads to another (the 'conclusions'). Typically, several verbally expressed propositions are combined so as to yield a single verbal conclusion, but many practical inferences depend on putting together information from the perceived state of affairs, general knowledge, and the verbal premises – the conclusion may never be formulated verbally but instead issue in a direct course of action. The process of *reasoning* depends on principles, either explicit or implicit, that establish some sort of relation between premises and conclusion. In the case of deduction, these principles are supposed to be logical and, in particular, to guarantee the validity of the inference, i.e. that its conclusion must be true given that its premises are true. There are other forms of reasoning depending on other underlying principles, but whenever an inference fails to conform to the principles of logic or goes beyond the premises in an unwarranted way, there is no guarantee that it is valid: the premises could be true, and the conclusion nevertheless false. For example, suppose you were to read in a paper:

> The victim was stabbed to death in a cinema during the afternoon.
> The suspect was travelling on a train to Edinburgh when the murder occurred.

You would probably conclude invalidly that the suspect was innocent. This is a typical piece of everyday reasoning, and it illustrates three important phenomena. Firstly, the inference depends both on the given premises and on many aspects of general knowledge, e.g. if one person stabs another they must be close to each other, people cannot be in two different places at the same time, and there are no cinemas on trains travelling to Edinburgh. Your use of such links in the inferential chain is so rapid and automatic that you are hardly aware of them. Indeed, the need for them was discovered only when people tried to devise computer programs that understood discourse[1]. Secondly, you drew a useful and informative conclusion of your own. Since an infinite number of logically valid conclusions (largely trivial) follow from any set of premises, you must have been guided by some principles entirely outside logic to reach your particular conclusion. Thirdly, although your conclusion is, strictly speaking, invalid, if it is challenged you can test its validity, i.e. whether it could be false given that the premises are true. When people are confronted with such a challenge, they search for alternatives and often produce ingenious scenarios in which the suspect is guilty, e.g. he (*sic*) may have had an accomplice; he may have left a spring-loaded knife in the cinema seat; he may have used a radio-controlled robot; and, fiendishly ingenious, he may have given the victim a post-hypnotic suggestion to stab himself[2].

There are many forms of reasoning, depending on the principles that are brought to bear on a problem, but these three phenomena of ordinary human reasoning stand most in need of explanation.

Logic and reasoning

Logic specifies the principles of valid reasoning (in certain domains), and a given branch of logic, such as the propositional calculus, i.e. the Boolean algebra of 'not', 'and', and 'or', can be formalized in many different ways. Most psychologists have assumed that there is a mental logic that enables us to reason[3–6], and the late Jean Piaget, the Swiss psychologist, argued that formal reasoning, which children are supposed to master in their early teens, is 'nothing more than the propositional calculus itself'[7]. According to this doctrine of mental logic, an inference is made by translating its premises into a mental language, adding the relevant pieces of general knowledge triggered during comprehension, and then applying formal rules of inference to these representations to derive a conclusion from them. The crucial questions are accordingly: what logic does the mind contain, and how is it represented there?

One of the many difficulties with Piaget's theory is that inferences which hinge on quantifiers, such as 'all' and 'some', cannot be captured within the propositional calculus. They call for

the more powerful quantificational calculus, which includes not only the propositional calculus but also an additional apparatus for quantifiers Computer programs for deductive reasoning have suggested a number of ways in which this logic might be represented in the mind, and I will outline the three most important approaches

Computer programs for reasoning

A major logical discovery of this century was that there is no decision procedure for the quantificational calculus There are algorithms that will determine in a finite amount of time that an inference is valid, but there can be no such procedure guaranteed to discover that an inference is invalid. A computer program must therefore minimize the time it takes to discover proofs, because as it grinds away there is no way of knowing whether it will ultimately yield a proof or merely go on computing for ever One economy is to use just a single rule of inference, the so-called 'resolution' rule

A or *B*
not-*A* or *C*
∴ *B* or *C*

which is a combination of two familiar disjunctive rules:

A or *B*	*A*
not-*A*	not-*A* or *C*
∴ *B*	∴ *C*

If not-*A* is the case, then *B* follows from the first of these rules; if *A* is the case, then *C* follows from the second of them. Since either *A* or not-*A* must hold, it follows that either *B* or *C* must, too Thus, whenever an assertion, *A*, and its negation, not-*A*, occur in disjunctive premises, they can be deleted to leave a disjunction of whatever is left behind The problem with the resolution rule is to get the premises into a form to which it can apply the rule only works if all quantifiers have been eliminated, and all the connectives ('and', 'if', etc) have been translated into equivalent inclusive disjunctions Table I summarizes the major steps of the method[8].

The resolution method is built into the programming language, PROLOG, which the Japanese have adopted in their quest for a 'fifth generation' of intelligent computers Yet, although the method is undoubtedly intelligent, it is highly artificial. It provides psychologists only with a standard of comparison, human beings are most unlikely to translate premises into a standard format and to use only a single rule of inference

The method of 'natural deduction' is altogether more plausible The logicians who devised this method intended it to be natural rather than parsimonious, and it is accordingly based on providing each connective and each quantifier with its own rules of inference Thus, there are rules for 'and', 'or', and 'if', e g

p	*p* or *q*	if *p* then *q*
q	not-*q*	*p*
∴ *p* and *q*	*p*	∴ *q*

TABLE I. A simple example of 'resolution' theorem-proving

The deduction to be evaluated
Premise 1 Fred is a scientist
Premise 2 All scientists are experimenters
Conclusion 3 Fred is an experimenter
Step 1 Translate the deduction into a *reductio ad absurdum* i.e negate the conclusion with the aim of showing that the resulting set of propositions is inconsistent
1 Scientist (Fred)
2 (For any *x*) (If Scientist(*x*) then Experimenter(*x*))
3 not(Experimenter(Fred))
Step 2 Eliminate the quantifier, 'any' its work can be done by the presence of variables (The quantifier 'some' is eliminated by representing it as a function, e g 'some' in 'all scientists have done some experiments' means if a scientist is provided, a value is delivered, namely, a set of experiments) Translate all connectives into equivalent inclusive disjunctions, e g 'if *A* then *B*' becomes 'not-*A* or *B*' Hence, assertion 2 becomes:
2 not(Scientist(*x*)) or Experimenter(*x*)
Step 3 Apply the Resolution rule to any premises containing inconsistent elements, e g assertion 3 is inconsistent with the occurrence of Experimenter(*x*) in 2 The rule entirely eliminates 3 and reduces 2, all that is left of the three assertions is
1 Scientist(Fred)
2 not(Scientist(*x*))
A second application of the rule eliminates these two assertions Whenever the set of assertions is reduced to nothing in this way, they are inconsistent the conclusion follows at once since we have obtained a *reductio ad absurdum* of its negation

Natural deduction is currently the most favoured hypothesis about the way in which logic is represented in the mind, and several theorists have formulated different sets of rules in order to try to capture differences in the difficulty of inferences[3,5,9,10].

Natural deduction, however, is just another way of formalizing inference the specific *semantic* content of the premises is irrelevant. Indeed, its irrelevance is the foundation on which all formal logic rests, since the set of valid deductions is to be captured solely in virtue of their form However, as we have already seen, the content of the premises is crucial because it triggers the additional knowledge that people bring to bear on a deduction. Hence, mental logic must be supplemented by some mechanism to handle this aspect of reasoning

One solution is to use only rules of inference with a specific content – a method that has been implemented in the programming language PLANNER and its descendants[1] PLANNER relies on the close resemblance between reasoning and planning an inference is a series of assertions each following from what has gone before and leading to a conclusion; a plan is a series of hypothetical actions each made possible by what has gone before and leading to a goal Hence, the derivation of a plan is organized in much the same way as the derivation of a proof Programs written in PLANNER-like languages have a data-base that consists of a set of assertions, written in a predicate-argument format, such as

(SCIENTIST FRED).

The assertion, 'Fred is a scientist', is accordingly true with respect to this data-base, and PLANNER allows a programmer to implement procedures that will evaluate a sentence with respect to the data-base and return its truth value, and that will take a new sentence and add the corresponding assertion to the data-base. However, if the new assertion is of the form

All scientists are experimenters

then rather than tamper with the data-base – going through it and long-windedly adding the relevant assertion about each scientist including Fred, PLANNER allows the result of such an operation to be described hypothetically in the following way A specific rule of inference is set up

TABLE II A simple inference using mental models

Step 1 Form a model based on the premises and any relevant general knowledge The premises
'Fred is a scientist'
All the scientists are experimenters'
are in fact self-sufficient The first premise yields the model
Fred = scientist
scientist
scientist
The content of the second premise is then added
Fred = scientist = experimenter
scientist = experimenter
scientist = experimenter
(experimenter)
Step 2 Find a relation in all the models so far constructed that is not stated explicitly in the premises, and formulate a conclusion expressing it
'Fred is an experimenter'
Step 3: Search for an alternative model of the premises that falsifies the putative conclusion If there is definitely no such model, the conclusion is valid, if no such model can be found but the search may not be exhaustive, the conclusion may be valid; if such a model is found, then return to step 2
The example does not depend on any information beyond the given premises, and since the model is finite it is easy to determine that any variant of it that severs the link between Fred and experimenter is no longer a model of the premises Hence, the conclusion is valid

(CONSEQUENT (X)
(EXPERIMENTER ?X)
(GOAL (SCIENTIST ?X)))

and what it says in effect is the program can infer the consequent that X is an experimenter, provided that it can establish the goal that X is a scientist. Hence, if the program is asked to evaluate the truth of the assertion, 'Fred is an experimenter', it searches the data-base for the corresponding assertion: (EXPERIMENTER FRED). If there is no such assertion there, it then looks to see whether it has any rules of inference in which the CONSEQUENT clause matches the input· (EXPERIMENTER FRED). The rule above, for instance, matches this assertion, and the value of the variable, ?X, now becomes, FRED The program's goal, as defined by the rule, is to establish.

(GOAL (SCIENTIST FRED))

and this time there is a corresponding assertion in the data base Since the goal is satisfied, the consequent is satisfied too, and the program can respond that the assertion, 'Fred is an experimenter' is true. In a realistic program, the rules of inference are likely to be rather more complicated and contain several goals. Likewise, there may be many rules that match an assertion or goal, and the program explores them in order to try to discover one that works

Goal-directed programs allow rules of inference to be formulated with a specific content every general assertion takes the form of such a rule, either a consequent rule of the sort illustrated here or some other sort of rule. These rules can even be supplemented by information specific to that content, such as heuristics about how to achieve a particular inferential goal, and they could readily trigger further factual information So-called 'expert systems', which are computer programs that provide advice on such matters as medical diagnosis, molecular structure, and drilling for minerals, are generally based on the same method of reasoning in which a specific goal is used to trigger a series of sub-goals However, as a model of human reasoning, goal-directed programs have one crucial defect· they provide absolutely no machinery for general inferential abilities They swing too far away from formal procedures. What is needed is the best of both worlds. general inferential ability combined with a sensitivity to content

Mental models in reasoning

An inference is valid if its conclusion cannot be false, given that its premises are true. One way in which a valid inference can be made is to imagine the situation described by the premises and then to consider if there is any way in which the conclusion could be false This method is semantic rather than syntactic as is the case with formal rules of inference. The reasoner builds a 'mental model' based on the premises and any relevant general knowledge, establishes a conclusion based on a relation in the model that was not stated in the premises, and then searches for alternative models of the premises that falsify the conclusion[2,11] What complicates this procedure is that there are usually many alternative situations which are compatible with the truth of the premises Given a premise, such as·

All the scientists are experimenters

how is one to build a single model that captures its content? The answer is to make some bold assumptions, which if need be can be revised later Thus, you can imagine that the set of scientists, which you know to exist and to be very large, consists of, say, just three token individuals

scientist
scientist
scientist

You may form a vivid image of these three individuals, but the theory assumes that what is crucial is not your subjective experience but the underlying structure of the model, which is not usually available to conscious introspection· a finite set of tokens represents a finite set of individuals Since the premise asserts that all the scientists are experimenters, you must add this information to your model

scientist = experimenter
scientist = experimenter
scientist = experimenter
(experimenter)

where the token in parentheses represents an individual who may or may not exist – an experimenter who is not a scientist, because the premise (and

TABLE III. An example of a 'three model' reasoning problem

The premises
None of the artists are beekeepers
All the beekeepers are chemists
The first model

a	
a	
———	where the barrier represents negation
b = *c*	
b = *c*	
(*c*)	

suggests the conclusion. None of the artists are chemists, or its converse, drawn erroneously by most subjects[12]
The second model

a	
a = (*c*)	
b = *c*	
b = *c*	

falsifies the conclusions above The two models together yield
Some of the artists are not chemists
Some of the chemists are not artists
The third model

a = (*c*)	
a = (*c*)	
b = *c*	
b = *c*	

eliminates the first of the previous pair of conclusions The valid conclusion is
Some of the chemists are not artists

your general knowledge) leave this possibility open Once you have constructed such a model of two or more premises, you can scan it for any interesting new relations to be used as the basis of a conclusion. When you have formulated a conclusion, you can check its validity by searching for alternative models of the premises that falsify it Table II summarizes the steps needed to make a simple valid inference using mental models

The main prediction of the mental model theory is obvious: the greater the number of alternative models that have to be constructed to draw a valid deduction, the harder the task will be. Here is an example of an inference that is much harder than the simple 'one model' problem in Table II. Suppose there are some artists, beekeepers and chemists in a room, and the following assertions are made about them:

None of the artists are beekeepers
All the beekeepers are chemists

What, if anything, follows validly from these premises? If readers wish to test their deductive ability, they should commit their answer to paper Relatively few intelligent adults respond correctly; the right answer for the right reasons calls for three mental models to be constructed, as is shown in Table III There are, of course, other possible forms of mental model, such as Euler circles and Venn diagrams, but only the number of mental models of the sort proposed by the present theory correctly predict the difficulty of a reasoning task[2,12]

The general algorithm for reasoning by mental models has been implemented in several computer programs It makes no use of formal rules of inference, but it is clearly necessary to specify procedures that construct appropriate models based on the meaning of the premises. This task is relatively easy for certain areas, e g spatial inferences, and inferences that depend only on the meanings of quantifiers and connectives[2] However, there is much that remains to be done in order to extend the theory to causal assertions and other domains of ordinary discourse outside the scope of formal logic, but well within the scope of everyday human reasoning

Selected references

1 Winograd, T (1972) *Understanding Natural Language*, Academic Press, New York
2 Johnson-Laird, P N (1983) *Mental Models Towards a Cognitive Science of Language, Inference, and Consciousness* Cambridge University Press, Cambridge, Harvard University Press, Cambridge, Mass
3 Braine, M D S (1978) *Psychol Rev* 85, 1–21
4 Osherson, D N (1975) in *Reasoning Representation and Process in Children and Adults* (Falmagne, R , ed), Erlbaum, Hillsdale, New Jersey
5 Rips, L J (1983) *Psychol Rev* 90, 38–71
6 Wason, P C and Johnson-Laird, P N (1972) in *The Psychology of Reasoning Structure and Content*, Batsford, London
7 Inhelder, B and Piaget, J (1958) *The Growth of Logical Thinking from Childhood to Adolescence*, Routledge and Kegan Paul, London
8 Robinson, J A (1965) *J Assoc Comput Mach* 12, 23–41
9 Braine, M D S and Rumain, B in *Carmichael's Manual of Child Psychology, 4th Edition Cognitive Development* (Flavell, J and Markman, E , eds), John Wiley and Sons, New York (in press)
10 Johnson-Laird, P N (1975) in *Reasoning Representation and Process in Children and Adults* (Falmagne, R , ed), Erlbaum, Hillsdale, New Jersey
11 Johnson-Laird, P N (1980) *Cognitive Sci* 4, 71–115
12 Johnson-Laird, P N and Bara, B (1984) *Cog* 16, 1–61

Philip N Johnson-Laird is Assistant Director of the M R C Applied Psychology Unit, Cambridge CB2 2EF, UK