Logic and Reasoning

PHILIP N. JOHNSON-LAIRD
Princeton University

GLOSSARY

deductive reasoning The process of establishing that a conclusion follows validly from premises (i.e., that it must be true given that the premises are true).

deontic reasoning Reasoning about actions that are obligatory, permissible, or impermissible.

formal rules of inference Rules that can be used to derive a conclusion from premises in a way that takes into account only the form, not the meaning, of the premises. Logical calculi rely on formal rules, and so do many psychological theories of reasoning.

implicit reasoning A fast, automatic, and largely unconscious process of making inferences in order to make sense of the world and of discourse (e.g., to select the appropriate sense of a word, or to establish the appropriate referent for a pronoun).

inductive reasoning The process of deriving plausible conclusions from premises.

logic The science of implications among sentences in a formalized language. Logical calculi are systems of proof based on formal rules of inference (proof theory); they have an accompanying semantics (or model theory).

mental models Representations of the world that are postulated to underlie human reasoning; each model represents what is true in a single possibility.

validity An inference is valid if its conclusion must be true given that its premises are true. A valid inference from true premises yields a true conclusion; a valid inference from false premises may yield a true or a false conclusion.

Logic captures the implications among sentences. A logical calculus consists of a precise definition of a language and a set of rules of inference that can be used to derive conclusions from premises. The rules are formal, that is, they operate on the form of sentences, not their meaning. The calculus, however, may have a semantics, which provides interpretations for all the sentences in the language. Modern logic lies at the heart of the development of computers and computer programming languages. However, logic is not easy to use in the evaluation of everyday inferences because no algorithm exists for translating such inferences into sentences in a logical calculus—a gap that the logician Bar-Hillel once referred to as the scandal of logic. Logic is also not a theory of how human beings reason. That topic is the province of psychology. Although psychologists studied deductive reasoning for almost the entire 20th century, they began to formulate theories of the process only in the past 25 years. Deductive reasoning is now under intensive investigation, and more is known about it than any other variety of thinking. The aim of this article is accordingly to outline the general principles of logic; to describe current theories of human reasoning, which owe much to logic; and to outline what is known about the role of the brain in reasoning.

I. LOGIC

From the founder of logic, Aristotle, onwards logicians have analyzed formal patterns of valid inference. A deduction is valid if its conclusion must be true given that its premises are true. The original aim of logic, as Leibniz remarked, was to replace rhetoric with calculation. Modern formal logic began during the last quarter of the 19th century, but nowadays logicians draw a sharp distinction between formal systems of logic, which they refer to as proof theory, and semantic systems of logic, which they refer to as model theory. The distinction is clearest in the case of the sentential calculus. This calculus concerns implications that depend on sentential negation, as expressed by "not," and various sentential connectives, such as "if," "and," and "or," which are treated in an idealized way. The following inference is an example of a valid deduction that can be proved in the sentential calculus:

> *If the brakes are on and the switches are on then the engine is ready to start.*
> *The brakes are on.*
> *The switches are on.*
> *Therefore, the engine is ready to start.*

The inference is based on three atomic sentences (i.e., sentences that contain neither negation nor any connectives): the brakes are on, the switches are on, and the engine is ready to start. The inference is valid and has the form

> *If A and B then C.*
> *A.*
> *B.*
> *Therefore, C.*

where A, B, and C, are variables that can take as values any sentences including those that in turn contain connectives.

Logicians can set up the proof theory for a calculus in various ways. They can formalize the sentential calculus, for example, using just a single rule of inference and a set of axioms, which are assertions that are assumed to be true. However, a more intuitive method, known as natural deduction, dispenses with axioms in favor of formal rules of inference for negation and for each of the sentential connectives. Certain rules introduce connectives into a proof, such as the rule that introduces "and," using it to conjoin two premises:

> *A.*
> *B.*
> *Therefore, A and B.*

Certain rules eliminate connectives from a proof, such as the well-known rule of modus ponens:

> *If A then B.*
> *A.*
> *Therefore, B.*

These two rules suffice to prove the conclusion about starting the engine:

1. If the brakes are on and the switches are on then the engine is ready to start.
2. The brakes are on.
3. The switches are on.
4. Therefore, the brakes are on and the switches are on [The rule for introducing "and" applied to sentences 2 and 3]
5. Therefore, the engine is ready to start. [Modus ponens applied to sentences 1 and 4].

Table I presents a set of formal rules of inference for the sentential calculus. With such rules, you can construct a formal proof, as in the preceding example, with each step in the proof warranted by one of the rules of inference.

Your knowledge of the meaning of the connectives helps you to understand the validity of the rules in Table I. However, the rules do not rely on these meanings. They work in a formal way, allowing you to write patterns of symbols given other patterns of symbols. A proof in a formal calculus is accordingly like a computer program. A computer predicts the weather, for example, but it has no idea of what rain or sunshine is or of what it is doing. It slavishly shifts "bits," which are symbols made up from patterns of electricity, from one memory store to another, and

Table I
Formal Rules of Inference for the Sentential Calculus[a]

A ∴ Not (Not A)	Not (Not A) ∴ A
A B ∴ A and B	A and B ∴ A
A ∴ A or B, or both	A or B, or both. Not A ∴ B
Rule for conditional proof A (a supposition) … B (i.e., B can be derived from A) ∴ If A then B	Rule for modus ponens If A then B A ∴ B

[a] The rules in the left-hand column introduce negation and the sentential connectives into inferences; those in the right-hand column eliminate them from inferences.

displays symbols that meteorologists can interpret as maps of weather. Indeed, proofs and computer programs are intimately related, and certain programs can prove inferences in logical calculi. Likewise, certain programming languages, such as PROLOG, are akin to a logical calculus.

Formal proofs establish that inferences are valid, but validity is not a concept that is defined within proof theory. Its definition hinges on truth, which underlies the semantics of the calculus (i.e., its model theory). In the model theory of the sentential calculus, the truth or falsity of compound sentences depends only on the truth or falsity of their constituent sentences. Thus, an assertion of the form, A or B or both, is true if A is true, B is true, or both of them are true. Otherwise it is false. Logicians lay out these definitions in truth tables, as shown in Table II. Each row in a truth table is a ''model'' of a possibility and presents the truth value of the compound sentence—in this case, A or B or both—in that possibility. The first row in the table, for instance, presents the case in which A is true and B is true, and so the disjunction is true in this possibility.

One problematic connective is ''if.'' Its everyday usage sometimes departs from its idealized logical meaning in the sentential calculus. An assertion such as

If that patient has malaria then she has a fever

is, in fact, compatible with three possibilities: The patient has malaria and a fever, she has no malaria and fever, and she has no malaria and no fever. It is false in only one case: She has malaria and no fever. The assertion is therefore equivalent to

If that patient has malaria then she has a fever, and if she does not have malaria then she either has or does not have a fever.

Logical license exists just as much as poetic license: Logicians make simplifying assumptions about the meanings of logical terms.

Table II
A Truth Table for the Disjunction *A or B or Both*, Which Shows Its Truth Value for the Four Possibilities Depending on the Truth or Falsity of A and of B

A	B	A or B or both
True	True	True
True	False	True
False	True	True
False	False	False

The validity of an inference in the sentential calculus can be established using the model theory of the calculus. Table III shows how premises can be used to eliminate possibilities from a truth table. When you have eliminated the impossible then, as Sherlock Holmes remarked, whatever remains, however improbable, must be the case. In other words, an inference is valid if the conjunction of its premises with the negation of its conclusion is inconsistent (i.e., not a single row in the resulting truth table contains the entry ''true''). For instance, if you conjoin the negation of the conclusion in Table III, ''The engine is not ready to start,'' to the premises, then it would eliminate the last remaining possibility in the truth table. It is therefore impossible for the premises to be true and for the conclusion to be false: The inference is a valid.

Any conclusion that can be proved using formal rules for the sentential calculus is also valid using truth tables and vice versa. There is also a decision procedure for the calculus; that is, the validity or invalidity of any inference can be established in a finite number of steps. Unfortunately, sentential inferences are computationally intractable. It is feasible to test the validity of inferences based on a small number of atomic sentences. However, as the number of atomic sentences in an inference increases, its evaluation in any system —no matter how large or how rapid—takes increasingly longer and depends on increasingly more memory, to the point that a decision will not emerge during the lifetime of the universe.

The sentential calculus has a decision procedure, but it is intractable. The predicate calculus includes the sentential calculus, but also deals with quantifiers—that is, with sentences containing such words as ''any'' and ''some,'' as in ''Any electrical circuit contains some source of current.'' The predicate calculus does not even have a decision procedure. Any valid inference can be proved in a finite number of steps, but no such guarantee exists for demonstrations of invalidity. Attempts to show that an inference is invalid may, in effect, get lost in the ''space'' of possible derivations. The principal discovery of 20th century logic, however, is Gödel's famous proof that no consistent calculus is powerful enough to yield derivations of all the valid theorems of arithmetic. Arithmetic is thus incomplete. This result drives a wedge between syntax (proof theory) and semantics (model theory). Any attempt to argue that semantics can be reduced to syntax is bound to fail. Semantics has to do with truth and validity, whereas syntax has to do with proofs and formal derivability.

Table III

The Validity of an Inference Is Shown Using a Truth Table[a]

1. If the brakes are on and the switches are on then the engine is ready to start.
2. The brakes are on.
3. The switches are on.

All that remains is the first possibility, and so it follows validly: Therefore, the engine is ready to start.

Brakes are on	Switches are on	The engine is ready to start	Possibilities that are eliminated
True	True	True	
True	True	False	Eliminated by 1
True	False	True	Eliminated by 3
True	False	False	Eliminated by 3
False	True	True	Eliminated by 2
False	True	False	Eliminated by 2
False	False	True	Eliminated by 2
False	False	False	Eliminated by 2

[a] The premises are used to eliminate possibilities.

II. DEDUCTIVE REASONING

Logic tells us about implications among sentences, but it is not a theory of human reasoning. This topic is a concern of psychology. In the last 25 years of the 20th century, psychologists proposed a variety of theories of reasoning—that it depends on a memory for previous cases, on rules that capture general knowledge, on ‘‘neural nets’’ representing concepts, or on specialized innate modules for matters that were important to our hunter–gatherer ancestors. However, humans have the ability to reason about matters for which they have no specific knowledge. Even if you know nothing about brakes, switches, and engines, you can grasp the validity of the earlier inference about them. This ability lies at the heart of the development of mathematics and logic. Hence, a critical question is whether it depends on syntactic or semantic principles. The following sections describe psychological theories of both sorts.

A. Formal Rule Theories

The first theories of human deductive ability postulated that the mind tacitly uses formal rules of inference like those of a system of natural deduction. Such theories continue to have many proponents, notably Daniel Osherson, Lance Rips, and the late Martin Braine and colleagues. Philosophers have proposed similar theories, and computer scientists have implemented formal systems for the computer generation of proofs. What these proposals have in common is the idea that reasoning depends on applying formal rules of inference to the premises of an inference in order to derive the conclusion in a sequence of steps akin to a proof.

Rips's PSYCOP theory was the first formal rule theory in psychology to cope with connectives and quantifiers and to be implemented in a computer program (written in PROLOG). The system is otherwise typical of formal rule theories. It postulates that reasoning depends on a single deterministic process, that it relies on natural deduction, and that it makes use of suppositions—sentences that are assumed provisionally for the sake of argument, and that have to be ‘‘discharged’’ if a proof is to yield a conclusion. There are two ways to discharge a supposition. First, it can be incorporated within a conditional conclusion (see the rule for conditional proof in Table I). Second, if a supposition leads to a contradiction, then it must be false given that the premises are true (according to the rule of ‘‘reductio ad absurdum,’’ which is not shown in Table I). As an example, consider the proof for an inference of a form known as *modus tollens*. There are two premises, such as

1. If the switches were not on then the engine did not start.

2. The engine did start.

The proof starts with a supposition:

3. Suppose: the switches were not on.
4. Therefore: the engine did not start. [Rule for modus ponens applied to 1 and 3]

There is now a contradiction between a sentence in the domain of the premises (The engine did start) and a sentence in the subdomain of the supposition (The engine did not start). The rule of reductio ad absurdum uses such a contradiction to negate, and thereby discharge, the supposition that led to the contradiction:

5. Therefore, the switches were on.

Like other formal rule theories, PSYCOP does not contain a rule for modus tollens, because the inference is more difficult for logically untrained individuals than modus ponens. Hence, it depends on the chain of inferential steps just given. In contrast, an inference of the following form, which we encountered earlier,

If A and B then C.
A.
B.
∴ *C.*

could be derived in two steps, first conjoining A and B, and then using modus ponens to derive the conclusion. However, the inference is so easy that PSYCOP has a single formal rule for drawing the inference (a conjunctive form of modus ponens).

Formal rule theorists try to postulate psychologically plausible rules of inference and a mechanism for using them to construct mental proofs. One problem is that unless certain rules, such as the rule for introducing "and" (see Table I), are constrained, they can lead to futile derivations:

The brakes are on.
The switches are on.
∴ *The brakes are on and the switches are on.*
∴ *The brakes are on and the brakes are on and the switches are on.*
∴ *The brakes are on and the brakes are on and the brakes are on and the switches are on.*

and so on ad infinitum. One solution is to incorporate the effects of such rules within other rules. In computer programs, however, a rule of inference can be used in two ways: either to derive a step in a chain of inference leading forward from the premises to the conclusion or to derive a step in a backward chain leading from the conclusion to a subgoal of proving its required premises. PSYCOP allows the dangerous rules to be used only in such backward chains, and thereby prevents them from yielding futile steps. PSYCOP therefore has three sorts of rules: those that it uses only forwards, such as the conjunctive rule for modus ponens; those that it uses only backwards, such as the rule for conditional proof; and those that it uses in either direction, such as the rule for modus ponens. A corollary is that reasoners should make modus tollens inferences only when they are given the putative conclusion, or when they can guess the conclusion and then try to prove it.

Given an inference to evaluate, PSYCOP always halts after a finite number of steps either with a proof of the conclusion or else in a state in which it has unsuccessfully tried all its possible derivations. Hence, the theory implies that people infer that a conclusion is invalid only if they fail to prove it. They carry out an exhaustive search of all possible derivations, and only then do they judge that the conclusion does not follow from the premises. However, valid inferences exist that PSYCOP cannot prove. If its exhaustive search has failed to find a proof, then there are two possibilities. Either the inference is invalid, or it is valid but beyond the competence of PSYCOP to prove. A psychological corollary is that people should never know for certain that an inference is invalid.

Formal rule theories postulate that the difficulty of a deduction depends on the number of steps in its derivation and the availability and ease of use of the required rules of inference. Modus ponens is easy because it depends on a single rule; modus tollens is more difficult because it depends on a chain of inferences. Formal rule theorists have corroborated their theories in experiments using large batteries of deductions. They estimate post hoc the probability of the correct use of each rule of inference. When these empirical estimates are combined appropriately for each inference, they yield a satisfactory fit with the difficulty of the inferences in the battery.

B. The Mental Model Theory

The mind may not contain any formal rules of inference unless an individual has learned logic. Instead, inferences could be based on an understanding of the meaning of the premises. Consider the following inference:

From where I stand, the peak of the mountain is directly behind the steeple. The old oak is on the

> *right of the steeple, and there is a flag pole between them. Therefore, if I move to my right so that the flag pole is between me and the peak of the mountain, the steeple is to the left of my line of sight.*

Reasoners might rely on axioms and formal rules to make this inference, but it seems more likely that they imagine the relevant spatial layout. This idea lies at the heart of the theory of mental models.

The theory postulates that mental models have three principal characteristics. First, each model represents a possibility. For example, the disjunction "The switches are on or the brakes are on, or both" calls for a separate model for each of the three possibilities (shown here on separate lines):

switches	
	brakes
switches	*brakes*

where "switches" denotes a model of the switches being on, "brakes" denotes a model of the brakes being on, and the third model combines the two.

Second is the principle of truth: Mental models represent only what is true and not what is false, and in this way they place a minimal load on working memory. Hence, the preceding models do not represent the row in the truth table in which the disjunction as a whole is false (Table II). Likewise, the first model represents that the switches are on, but it does not represent explicitly that in this possibility it is false that the brakes are on. People make a mental "footnote" about what is false, but normally they soon forget it. If they retain such footnotes, however, then they may be able to flesh out their mental models to make them fully explicit. Table IV presents the mental models and the fully explicit models for sentences based on each of the main sentential connectives. Mental models are accordingly like truth tables in which there are no "false" entries.

Third, the structure of a model corresponds to the structure of the situation that the model represents. A model is accordingly like a biologist's model of a molecule. The previous notation for the models fails to capture their rich internal structure. Visual images can be derived from some models, but models are often not visualizable. Early formulations of the theory concerned only the logical terms in the language, but recently the theory has been extended to deal with various sorts of nonlogical terms, such as spatial and temporal relations, and general knowledge about causal relations.

Table IV

Models for the Sentential Connectives[a]

Connective	Mental models		Fully explicit models	
A and B	A	B	A	B
A or else B	A		A	¬B
		B	A	B
A or B or both	A		A	¬B
		B	¬A	B
	A	B	A	B
If A then B	A	B	A	B
	…		¬A	B
			¬A	¬B
If and only if A then B	A	B	A	B
	…		¬A	¬B

[a] The middle column shows the mental models postulated for human reasoners, and the right-hand column shows fully explicit models, which represent the false components in true possibilities using negations that are true: "¬" denotes negation and "…" denotes a wholly implicit model. The footnote on the mental models for "if" indicates that the implicit model represents the possibilities in which A is false, and the footnote on the mental models for "if and only if" indicates that the implicit model represents the possibilities in which both A and B are false.

Reasoners use all the information available to them to construct models—discourse, perception, general knowledge, memory, and imagination. They formulate a conclusion that holds in their models but that was not explicit in the starting information. If a conclusion holds in all the models of the premises, then it is necessary given the premises. If it holds in at least one model of the premises, then it is possible given the premises. The probability of a conclusion depends on the proportion of models in which it holds, granted that each model is equiprobable, which is an assumption that reasoners make in default of evidence to the contrary. The theory accordingly unifies reasoning about necessity, possibility, and probability. They all depend on a semantic process rather than a formal one. They all depend on a grasp of meaning, which is used to imagine the possibilities compatible with the premises.

To illustrate the theory, consider the following inference:

The switches are on or the brakes are on, or both.
The switches are not on.
∴ *The brakes are on.*

The disjunctive premise elicits the models:

switches	
	brakes
switches	*brakes*

The second premise eliminates the models representing the possibilities in which the switches are on. The remaining model yields the conclusion that the brakes are on. This conclusion is valid because it holds in all the models—in this case, the single model—of the premises.

C. Five Empirical Phenomena

Psychological investigations have established five principal phenomena of deductive reasoning. The first phenomenon is that the more possibilities that reasoners have to envisage to draw an inference, the more difficult the inference—it takes them longer, and they are more likely to make a mistake. A simple example is that inferences based on a disjunction are more difficult when the disjunction is inclusive, as in the preceding example, than when it is exclusive and allows only two possibilities: "The switches are on or the brakes are on, but not both." The same effect of number of possibilities occurs in reasoning with other sentential connectives, in reasoning about spatial and temporal relations, and in reasoning with premises containing quantifiers, such as "all," "some," and "none."

The second phenomenon is that reasoners use counterexamples to establish invalidity. When reasoners draw conclusions for themselves, they may not consider counterexamples. However, when they reject a conclusion, they can do so by constructing a counterexample—that is, they envisage a possibility that satisfies the premises but refutes the conclusion. One experiment, for example, used problems, such as

> *More than half of the people in the room speak French.*
> *More than half of the people in the room speak English.*
> *Does it follow that more than half of the people in the room speak both French and English?*

Most people responded correctly, "no," and they typically reported having envisaged a situation analogous to the one represented in Fig. 1. They drew such diagrams when they were allowed paper and pencil. They also used counterexamples when they manipulated external models—cut-out paper shapes—in order to reason with quantifiers.

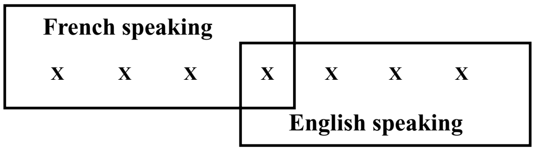

Figure 1 A counterexample used to refute an inference. Each x represents an individual: more than half of them speak French, and more than half of them speak English, but it is false that more than half speak both languages.

The third phenomenon is that human reasoners spontaneously develop a variety of different strategies in deductive reasoning. They do not use a single deterministic strategy. For example, in reasoning based on multiple premises containing sentential connectives, some individuals develop the strategy of translating each disjunctive premise into a conditional, some base their inferences on the most informative premise, and some make use of suppositions—even when there are categorical assertions among the premises. Many distinct inferential strategies occur, but the space of possible strategies has yet to be mapped.

The fourth phenomenon is the occurrence of illusory inferences. These inferences are compelling but invalid. The following is a typical example:

> *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?*

Most people respond "yes." The first premise is compatible with the possibilities:

King	
	Ace
King	*Ace*

They support the conclusion that an ace is possible. The second premise supports the same conclusion, and so reasoners are likely to respond affirmatively. However, this response overlooks the fact that when one premise is true, the others are false. Thus, if the first premise is true, the second premise is false. In which case, there cannot be an ace. Indeed, if there were an ace in the hand, then the first two of the premises would be true, contrary to the rubric that only one of the premises is true.

The rubric "only one of the premises is true" is equivalent to an exclusive disjunction, and a compelling illusion occurs in the following inference about a particular hand of cards:

If there is a king in the hand then there is an ace in the hand, or else if there isn't a king in the hand then there is an ace in the hand.
There is a king in the hand.
What, if anything, follows?

Nearly everyone, experts and novices alike, infers that there is an ace in the hand. It follows from the possibilities that people envisage. However, given the disjunction of the two conditionals, it is an error. The disjunction implies that one or other of the conditionals could be false. If the first conditional is false, then even the presence of a king fails to guarantee that there is an ace in the hand. The fallacies arise from a failure to think about what is false. It follows that any manipulation that emphasizes falsity should alleviate them. This prediction has been corroborated experimentally.

The fifth phenomenon is that knowledge and beliefs affect both the interpretation of premises and the process of reasoning. Consider, for example, the following conditional assertion:

If she played a sport then she didn't play soccer. Conditionals are normally compatible with three possibilities (see the fully explicit models in Table IV):

sport ¬ soccer
¬ sport ¬ soccer
¬ sport soccer

where ¬ denotes negation. However, the meaning of the noun soccer entails that it is a sport, and so knowledge of this meaning automatically rules out the third of these possibilities. General knowledge and knowledge of the context of an utterance can also eliminate possibilities. Individuals often know what the different possibilities are, and such knowledge modulates the interpretation of assertions. As an illustration, consider the following conditional:

If you strike a match properly then it lights.

Its interpretation includes the salient possibility:

strike lights

As often happens in discourse, however, the antecedent of the conditional fails to describe in complete detail the context in which the consequent holds. There are many circumstances in which a match will not light even if you strike it properly. You know, for instance, that if it is soaking wet it will not light. In fact, you have knowledge of the following explicit possibilities:

soak ¬lights
¬ soak ¬lights
soak lights

Now, suppose you soak a match in water and then strike it. What happens? The conditional implies that it lights. Your knowledge implies that it does not light. Your knowledge, however, takes precedence over the possibilities that the conditional asserts.

Given the following premises in a form known as a syllogism,

All the Frenchmen are wine drinkers.
Some of the wine drinkers are gourmets.

the majority of reasoners draw the plausible conclusion:

Some of the Frenchmen are gourmets.

However, with the next premises, which are identical in form,

All the Frenchmen are wine drinkers.
Some of the wine drinkers are Spanish.

few reasoners draw the conclusion

Some of the Frenchmen are Spanish.

They envisage the possibility in which the wine drinkers are of both nationalities, but they search more assiduously—and successfully—for a counterexample because this conclusion is preposterous. Hence, the main effect of beliefs on the process of reasoning is that they influence invalid inferences far more than valid inferences: People refrain from drawing unbelievable invalid conclusions.

The difficulty of coping with falsity and the effects of content come together in a well-known reasoning problem, Wason's selection task, which has been studied experimentally more than any other paradigm of reasoning. Table V presents two versions of the task, one with a neutral conditional and one with a deontic conditional concerning what is permissible. The difficulty of the version with the neutral conditional, "If a card has an 'A' on one side then it has a '2' on its other

Table V

Two Examples of Wason's Selection Task

A	B	2	3

1. The participants know that each card above has a letter on one side and a number on the other side. Their task is to select those cards that they need to turn over to discover whether the following conditional is true or false about the four cards:

If a card has an "A" on one side then it has a "2" on its other side.

Most people correctly select the "A" card, and some select the innocuous "2" card too. They fail to select the 3 card. However, if it has an A on its other side, the conditional is false.

Drinking	Not drinking	21 years	16 years

2. The participants know that each card above represents a person. One side states whether or not the person is drinking alcohol, and the other side states the age of the person. The task is to select those cards that need to be turned over to discover whether or not a person is violating the following conditional rule:

If persons are drinking then they are over the age of 20 years.

Most people correctly select the "Drinking" card and the "16 years" card.

side," arises from the participants' inability to base their selections on the possibility that falsifies the conditional:

$A \quad \neg 2$

They need to choose those cards that could be instances of this case, i.e., A and 3 (which is an instance of ¬ 2). With neutral conditionals, reasoners appear merely to select cards on the basis of their mental models of the conditional rather than its falsifying instance.

When the selection task concerns what is permissible or impermissible, such as breaking a social contract, then reasoners tend to make the correct selections (Table V, problem 2). Some psychologists argue that this version of the task maps onto mental schemas with a content that concerns such deontic matters. Evolutionary psychologists propose that social contracts mattered to our hunter–gatherer ancestors, and so an innate module evolved for reasoning about cheaters. What appears to be the case, however, is that any experimental manipulation that helps reasoners to envisage false instances of conditionals improves their performance in the selection task. Knowledge of cheating is just such a cue, but there are others. Experiments have shown, for example, that instructions to check for violations or to envisage counterexamples improve performance. The context of a conditional can also exert such effects. One study enhanced the participants' selections with a neutral conditional, "If A then 2." The participants were told that it was a rule followed by a machine that prints cards. The machine went wrong, and now the participants must check that it is printing out cards correctly.

Reasoners are sensitive to the likelihood of encountering potential counterexamples, and so some theorists have introduced probabilistic considerations into their analyses of the selection task. They defend a normative approach of this sort, arguing that participants rationally seek to maximize the expected gain in information from selecting a card. If they were testing in the real world, the following conditional:

If a creature is a raven then it is black.

it would make sense to examine creatures that are black because there are many fewer black than nonblack creatures. Hence, the argument goes, people are rational in selecting 2 rather than 3 to test the neutral conditional. In one study, however, participants were each paid 1000 pesetas (about $7) before carrying out the selection task with a neutral conditional. They were charged 250 pesetas for each card that they selected, but they were told that they could keep whatever money they did not spend provided that their evaluation of the conditional was correct. This incentive failed to improve performance. Likewise, individuals with higher SAT scores tend to do better on the selection task than those with lower scores.

Whatever "rational" is taken to mean, it seems inappropriate to apply it to those who lose money rather than gain it and to those who score lower on tests of cognitive ability.

The five sorts of phenomena reviewed in this section were all predicted by the model theory, although readers will need to consult the literature for the derivation of the predictions. Formal rule theories allow that knowledge and beliefs can affect the interpretation of premises; otherwise, the phenomena are difficult for these theories to accommodate.

III. IMPLICIT INFERENCES

Psychologists distinguish between the deliberative thinking that underlies deduction and the implicit, automatic, and largely unconscious inferences that help people to make sense of the world and its descriptions. Consider, for example, the following passage:

> *The pilot put the plane into a stall just before landing on the strip. He just got it out of it in time. It was a fluke.*

Readers have no difficulty in understanding the passage, but every noun and verb in the first sentence is ambiguous. Also the search for the referents for the three occurrences of the pronoun "it" in the passage defeats even the most advanced computer programs for interpreting natural language. Humans have no difficulty with the passage because they are equipped with a powerful system that uses general knowledge to make implicit inferences. Readers should also have no difficulty in understanding the following passage:

> *Apart from her husband, a hairdresser, Eve was the only woman among 52 men on the tour. As a costumier, she filled a much needed gap, because when a company of actors is putting on a play in a different town each night, no damage to the costumes is too trivial not to be mended.*

In fact, most people do not notice that the passage contains three deliberate mistakes. It implies that Eve's husband is a woman. It states that what is needed is a gap rather than Eve. It also asserts that no damage to the costumes is too trivial not be mended instead of what it surely means—no damage to the costumes is too trivial to be mended. The system of implicit inferences overrides the literal interpretation of the sentences and makes sense out of nonsense. The inferences resolve the senses of words and determine the references of pronouns and other such expressions. They enable individuals to construct a single model of the situation described in a passage, and the implicit system does not attempt to search for alternative models unless it encounters evidence for them. The process is therefore rapid, and it becomes as automatic as any other cognitive skill that calls for no more than a single mental representation at a time. For the same reason, implicit inferences lack the guarantee that their conclusions are valid. They are inductions rather than deductions. However, the implicit system is not isolated from the mechanisms of deduction. Normally, the two systems work together in tandem.

One consequence of implicit inferences is that people often jump to a conclusion, which later they have to withdraw. In logic, if a conclusion follows validly from premises, then no additional premises can invalidate it. Logic means never having to be sorry about a conclusion. As new premises are added to existing premises, then increasing numbers of logical conclusions follow (i.e., logic is "monotonic"). However, in daily life, conclusions are often withdrawn in the light of subsequent information. These inferences are "nonmonotonic". The original conclusion may have been based on an assumption made by default that turned out to be false. For instance, I tell you about my cat Hodge, and from your knowledge of cats you infer that Hodge has fur and a tail. You withdraw your conclusion, however, when you learn that Hodge is bald and tailless. Your knowledge contains various assumptions that you can make in default of information to the contrary. The whole purpose of these default assumptions is to allow you to make useful inferences that you can withdraw in the light of contrary evidence.

A more problematic sort of nonmonotonic reasoning is illustrated in the following example. You believe the following premises:

> *If Viv has gone shopping then she will be back in an hour.*
> *Viv has gone shopping.*

It follows, of course, that Viv will be back in an hour. However, suppose that Viv is not back in an hour. You are in a typical everyday situation in which there is a conflict between the consequences of your beliefs and the facts. At the very least, you have to withdraw your conclusion. You also have to modify your beliefs, but in what way? Should you cease to believe that Viv went shopping or that if she went shopping she will be back in an hour, or both? Philosophers and students of artificial intelligence have made various proposals about these puzzles. Unfortunately, the understanding of nonmonotonicity in human reasoning lags behind.

Reasoning in daily life often calls for the generation of explanations and diagnoses. For example, in the case of Viv's failure to return, you do not merely modify your beliefs, you try to make diagnostic inferences about what happened:

Possibly, Viv met a friend and went for a coffee.
Possibly, Viv felt ill on the way to the shops.

One possibility leads in turn to further explanatory possibilities, for example,

Possibly, Viv couldn't get the car to start after shopping.
∴*Possibly, the car's battery is dead. Possibly, Viv left the headlights on.*

You use your knowledge and any relevant evidence to generate possibilities. Human reasoners easily outperform any current computer program in envisaging putative explanations. Given two sentences selected at random from different stories, such as

Celia made her way to a shop that sold TV sets.
She had recently had her ears pierced.

they readily offer such explanations as Celia was getting reception in her ears and wanted the TV shop to investigate, or Celia had bought some new earrings and wanted to see how they looked on closed-circuit TV. This propensity to generate explanations underlies both science and superstition. The difference is that scientists test their explanations empirically.

Inferences in real life are often not deductively "closed"—that is, there is not enough information to draw a valid conclusion. Reasoners must therefore make inductions, that is, they use their knowledge to draw conclusions that go beyond the information given and that therefore may be false. There is no normative theory of induction and no comprehensive psychological theory of it, either. What does exist are a number of well-established heuristics, which were identified by two pioneers, Kahneman and Tversky. One heuristic is the availability of relevant knowledge. Most individuals, for example, judge that more people die in automobile accidents than as a result of stomach cancer. They are wrong, but the media publish more stories about auto accidents than about stomach cancer. Similarly, people rely on the representativeness of evidence. If you are told that Bill is intelligent but unimaginative and lifeless, then you are unlikely to judge that he plays jazz for a hobby, though you may find it more likely that he is an accountant who plays jazz for a hobby. If so, you have violated the principle that a conjunction (being an accountant and playing jazz) cannot be more probable than one of its components (playing jazz). The description of Bill, however, is more representative of an accountant than of a jazz musician. It has therefore led you to overlook a simple principle of probability.

IV. REASONING AND THE BRAIN

The famous Russian neuropsychologist Luria once remarked, "The cerebral organization of thinking has no history whatsoever." Fodor, the distinguished philosopher of mind, predicted that it has no future either because thinking depends on general processes rather than separate brain modules, such as those that underlie perception or motor control. Nevertheless, a start has been made in the study of the neuropsychology of reasoning. The results so far have been largely at the level of "these areas of the brain underlie reasoning," and their interpretations are at best tentative.

A. Logical Reasoning and Personal Reasoning

Clinical studies in the early 20th century often reported the loss of "abstract thinking" as a result of brain damage. Such accounts, however, suffered from two irremediable problems. On the one hand, they never succeeded in characterizing a principled difference between abstract and concrete thinking. On the other hand, they failed to pin down the particular effects of lesions in different parts of the brain. This shortcoming is understandable given that many regions of the brain are likely to underlie reasoning. Modern neuropsychological investigations suggest that the real distinction is between logical reasoning with neutral materials and personal reasoning that engages individuals' beliefs and knowledge (Table V). Some studies suggest that logical reasoning depends on the left cerebral hemisphere, whereas personal reasoning implicates the right hemisphere and bilateral ventromedial frontal cortex. Positron emission tomography scans show greater left hemisphere activity when individuals evaluate syllogisms, such as

All men have sisters.
Socrates was a man.
∴ Socrates had a sister.

or judge the plausibility of inductive inferences, such as

Socrates was a great man.
Socrates had a wife.
∴ All great men have wives.

The control task was to judge how many of the sentences had people as their subjects. The effects of brain damage also appear to support the dissociation between logical and personal reasoning. For example, left hemisphere lesions impair simple relational inferences, such as

Mary is taller than John.
John is taller than Anne
Is Mary taller than Anne?

People who live in nonliterate cultures are happy to carry out personal reasoning, but they balk at logical reasoning when the content is outside their experience. Analogous effects have been obtained using electroconvulsive therapy (ECT), which suppresses cortical activity for 30 min or more. Before ECT, the patients (depressives and schizophrenics) tended to justify their responses to deductive problems on logical grounds. They also did so more rapidly and confidently after ECT had suppressed their right hemispheres. However, after the suppression of their left hemispheres, they tended to respond on grounds of personal experience in ways similar to members of nonliterate cultures, often rejecting a logical task based on unfamiliar content as impossible because it was outside their knowledge. Similar effects of brain damage occurred in a study of the selection task with a neutral conditional (Table V). Patients with left hemisphere damage, like control subjects, tended to err in the characteristic way. Surprisingly, however, half the patients with right hemisphere damage made the correct selections.

Perhaps the right hemisphere impedes logical reasoning because it allows knowledge and probabilistic considerations to influence performance. Certainly, the right hemisphere seems to play a role in automatic implicit inferences. Given the passage,

Sally approached the movie star with pen and paper in hand. She was writing an article about famous people's views about nuclear power.

normal individuals are likely to infer that Sally wanted to ask the star about nuclear power. Patients with damage to the right hemisphere infer that Sally wanted the movie star's autograph. They are misled by the first sentence and cannot make the implicit inference from the second sentence to revise their interpretation. Patients who have had a right-hemisphere lobectomy are also poorer at reasoning from false premises than those with a left hemisphere lobectomy. In general, right hemisphere damage seems to impair the ability to "get the point" of a story, to make implicit inferences establishing coherence, and to grasp the force of indirect illocutions such as requests framed in the form of questions.

It is tempting, but erroneous, to conclude that the left hemisphere is the seat of logic, whereas the right hemisphere is the seat of personal reasoning. Damage to the right hemisphere can lead to semantic difficulties in the interpretation of words, and so it may also impair the comprehension of discourse. For instance, it impairs the deduction of converse relations, such as

John is taller than Bill.
Who is shorter?

A recent functional magnetic resonance imaging (fMRI) study confirmed the existence of dissociable networks for logical and personal reasoning, which share circuits in common in the basal ganglia, cerebellum, and left prefrontal cortex. However, the activation suggested that personal reasoning recruits the left hemisphere linguistic system, whereas logical reasoning—even in inferences of an identical form—recruits the parietal spatial system. Also, when reasoning elicits a conflict between logic and belief, right prefrontal cortex becomes active, perhaps to resolve the incongruency. Another recent fMRI study established that deductive reasoning activates right dorsolateral prefrontal cortex whereas mental arithmetic from the same premises does not. This study also showed that when an inference depends on a search for a counterexample then the right frontal pole is activated.

Frontal cortex plays a crucial role in decision making, as shown in a major series of studies carried out by Damasio and colleagues. They also investigated the selection task in testing the consequences of their somatic marker hypothesis. This hypothesis postulates that ventromedial frontal cortex underlies the typical "gut reaction" on which implicit everyday decisions rely. Considerable evidence supports this hypothesis: For example, individuals with frontal lesions tend to go bankrupt in real life and in laboratory gambling tasks. Similarly, the investigators found that patients with lesions in ventromedial frontal cortex were unaffected by whether the selection task was based on familiar or unfamiliar neutral contents. However, patients with lesions in other areas, like normal individuals, showed the characteristic effects of content. Correct performance in the selection task depends on grasping what counts as a counterexample to the conditional assertion.

B. Imagery and Spatial Representations

Does deductive reasoning rely on visual imagery? Behavioral studies have produced little evidence to suggest this is the case. Readers might suppose that this lack of evidence counts against the model theory. This view, however, confuses models with images. The model theory distinguishes between the two: Mental models are structural analogs of the world, whereas visual images are the perceptual correlates of certain sorts of model from a particular point of view. Indeed, many mental models are incapable of supporting visual images because they represent properties or relations that are not visualizable, such as ownership, obligation, and possibility. Recent studies have sharpened the need to distinguish between the degree to which relations evoke spatial models as opposed to visual images. The studies examined three sorts of materials, as rated by an independent panel of judges:

1. Relations that are easy to envisage spatially and easy to visualize, such as above, below, in front of, and in back of
2. Relations that are not easy to envisage spatially but are easy to visualize, such as cleaner, dirtier, fatter, and thinner
3. Control relations that are neither easy to envisage spatially nor easy to visualize, such as better, worse, smarter, and dumber

The studies examined both conditional inferences and inferences about simple relations among entities. They showed that inferences were faster with contents that were easier to envisage spatially than with the control contents, which in turn were faster than contents that were easy to visualize but difficult to envisage spatially. It seems that a relation such as "dirtier", elicits a visual image, but one that is irrelevant to the construction of a mental model that allows reasoners to make the required inference. In contrast, a relation, such as "in front of" elicits a spatial model that helps individuals to draw the inference. An fMRI study has also examined spatial reasoning. Given spatial problems, such as

> *The red rectangle is in front of the green rectangle.*
> *The green rectangle is in front of the blue rectangle.*
> *Does it follow that the red rectangle is in front of the blue rectangle?*

significant activation occurred in regions of parietal cortex that are known to represent and to process spatial information. Moreover, there was no reliable difference in the degree of activation between the right and the left hemispheres. Clinical studies of how brain damage affects the use of imagery in reasoning have produced mixed results, perhaps because they have not separated the two sorts of contents—spatial and nonspatial—that are both easy to visualize.

In summary, clinical and imaging studies of the brain have yet to establish how reasoners make deductions. There is evidence for separate systems mediating logical inferences with neutral content and personal inferences with a content that engages knowledge and beliefs. Future studies may determine whether separate brain mechanisms underlie the control of different deductive strategies, the use of diagrams as opposed to verbal premises, and the construction and evaluation of multiple models.

V. CONCLUSIONS

Modern logic has developed both proof theory and model theory for systems powerful enough to cope with all the deductive inferences that human beings make. What is lacking is a systematic method for translating such inferences into formal logic. Psychologists continue to investigate deductive reasoning. Their two main theoretical accounts are based on rules of inference and on mental models, respectively—a distinction that parallels the one between proof theory and model theory in logic. Rule theorists emphasize the automatic nature of simple deductions and postulate rules corresponding to them. More complex inferences, they assume, call for sequences of simple deductions. In contrast, model theorists emphasize that reasoning is the continuation of comprehension by other means. The system for implicit inferences based on knowledge aids the process of constructing models of discourse. In deliberative reasoning, individuals tend to focus on possibilities in which the premises are true. However, they can grasp the force of counterexamples. The evidence suggests that people have a modicum of deductive competence based on mental models. Rules of inference and mental models, however, are not incompatible. Advanced reasoners may construct formal rules for themselves—a process that ultimately leads to the discipline of logic.

See Also the Following Articles

ARTIFICIAL INTELLIGENCE • CATEGORIZATION • CREATIVITY • INFORMATION PROCESSING • INTELLIGENCE • LANGUAGE AND LEXICAL PROCESSING • PROBLEM SOLVING

Suggested Reading

Baron, J. (1994). *Thinking and Deciding*, 2nd ed. Cambridge Univ. Press, New York.

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

Brewka, G., Dix, J., and Konolige, K. (1997). *Nonmonotonic Reasoning: An Overview*. CLSI Stanford Univ. Press, Stanford, CA.

Evans, J. St. B. T., and Over, D. E. (1996). *Rationality and Reasoning*. Psychology Press, Hove, UK.

Garnham, A., and Oakhill, J. (1994). *Thinking and Reasoning*. Blackwell, Cambridge, MA.

Jeffrey, R. (1981). *Formal Logic: Its Scope and Limits*, 2nd ed. McGraw-Hill, New York.

Johnson-Laird, P. N. (2001). Mental Models and deduction. *Trends in Cognitive Scien.* **5,** 434–442.

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

Johnson-Laird, P. N., Legrenzi, P., Girotto, V., Legrenzi, M., and Caverni, J.-P. (1999). Naive probability: A mental model theory of extensional reasoning. *Psychol. Rev.* **106,** 62–88.

Kahneman, D., Slovic, P., and Tversky, A. (Eds.) (1982). *Judgment under Uncertainty: Heuristics and Biases*. Cambridge Univ. Press, New York.

Oaksford, M., and Chater, N. (1998). *Rationality in an Uncertain World: Essays on the Cognitive Science of Human Reasoning*. Psychology Press, Hove, UK.

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

Schaeken, W., De Vooght, G., Vandierendonck, A., and d'Ydewalle, G. (2000). *Deductive Reasoning and Strategies*. Erlbaum, Mahwah, NJ.

Stanovich, K. E. (1999). *Who Is Rational? Studies of Individual Differences in Reasoning*. Erlbaum, Mahwah, NJ.

Wharton, C. M., and Grafman, J. (1998). Deductive reasoning and the brain. *Trends Cognitive Sci.* **2,** 54–59.