Opinion

Deductive reasoning

Phil Johnson-Laird*

This article begins with an account of logic, and of how logicians formulate formal rules of inference for the sentential calculus, which hinges on analogs of negation and the connectives *if, or*, and *and*. It considers the various ways in which computer scientists have written programs to prove the validity of inferences in this and other domains. Finally, it outlines the principal psychological theories of how human reasoners carry out deductions. © 2009 John Wiley & Sons, Ltd. *WIREs Cogn Sci* 2010 1 8–17

Deductive reasoning is the mental process of making inferences that are logical. It is just one sort of reasoning. But, it is a central cognitive process and a major component of intelligence, and so tests of intelligence include problems in deductive reasoning. Individuals of higher intelligence are more accurate in making deductions,[1] which are at the core of rationality. You know, for instance, that if your printer is to work then it has to have ink in its cartridges, and suppose that you discover that that there is no ink in its cartridges. You infer that the printer would not work. This inference has the important property of logical validity: if its premises are true then its conclusion must be true too. Logicians define a *valid* deduction as one whose conclusion is true in every possibility in which all its premises are true (Ref 2, p. 1). All able-minded individuals recognize that certain inferences are valid because there are no counterexamples to them, that is, no possibilities in which the premises hold but the conclusion does not. This idea underlies deductive reasoning. And deductive reasoning in turn underlies the development of all intellectual disciplines and our ability to cope with daily life. The topic is studied in logic, in artificial intelligence, and in cognitive science. Hence, the aim of this interdisciplinary review is to survey what these different disciplines have to say about deduction, and to try to solve the mystery of how individuals who know nothing of logic are nevertheless able to reason deductively.

The plan of the article is straightforward. It starts with logic, because logic began as a systematic attempt to evaluate inferences as valid or invalid, and because a knowledge of logic informs our understanding of both computer programs for deduction and theories of human deduction. The article next reviews computer systems for deductive reasoning. Its final section considers human reasoning, and outlines the principal attempts to make sense of it.

LOGIC

A test case throughout this article is sentential deduction, which hinges on negation and the connectives (in English): *if, and*, and *or*. A logical calculus for sentential reasoning has three main components.[2] The first component is a grammar that specifies all and only the well-formed sentences of the language. The sentential calculus is not concerned with an analysis of the internal structure of simple *atomic* sentences, such as: ‘There is a circle’, which it merely assigns to be the values of variables, such as a, b, and c. Hence, a compound sentence, such as ‘There is a circle and there is not a triangle’, is represented as: $a \,\&\, \neg\, b$, where ‘&’ denotes logical conjunction, ‘¬’ denotes negation, and ‘a’ and ‘b’ are variables whose values are the appropriate atomic sentences. In a simple version of the calculus, there are just two other connectives: ‘v’, which denotes an inclusive disjunction equivalent to: *a or b or both*, and ‘ →’, which denotes the closest analog in logic to the conditional assertions of daily life. For example, ‘if there is a circle then there is a triangle’ is represented as: $a \rightarrow b$. Conditionals in daily life can have many interpretations,[3] and so to avoid confusion logicians refer to ‘ →’ as ‘material implication’.

The grammar of the sentential calculus is simple. It has variables (a, b, c, etc.), negation (¬), three connectives (&, v, and →), and three rules for forming sentences:

$$\text{sentence} = \text{variable}$$

$$\text{sentence} = \neg\ (\text{sentence})$$

$$\text{sentence} = (\text{sentence connective sentence})$$

*Correspondence to: phil@princeton.edu

Department of Psychology, University of Princeton, Princeton, NJ 08540, USA

DOI: 10.1002/wcs.20

TABLE 1 | Examples of 'natural deduction' rules of inference of the sort adopted in psychological theories based on formal logic.

Formal rules for introducing connectives:		
A		
B	A	
∴ A & B	∴ A v B	
Formal rules for eliminating connectives:		
	A v B	A → B
A & B	¬ A	A
∴ B	∴ B	∴ B

Formal rules for introducing connectives, where '*A* |–*B*' signifies that the supposition of *A* for the sake of argument yields with other premises a proof of *B*.)

This grammar specifies that each of the following examples is a sentence in the logic, where brackets are omitted to simplify matters:

a
¬*b*
(*a* → ¬ *b*) v ¬ (*c* & *d*)

The second component of the calculus is a set of rules of inference that enable proofs to be derived in a purely formal way. In fact, there are many ways to couch such rules. One way that seems intuitive is a so-called 'natural deduction' system[4] that has rules for introducing connectives and rules for eliminating them. For example, conjunction has a rule that introduces '&' by using it to combine any two premises, which themselves may be compound:

A
B
Therefore, *A* & *B*

Another rule eliminates '&' by drawing a conclusion corresponding to one of the sentences that it conjoins:

A & *B*
Therefore, *A*

A similar rule allows *B* to be derived from *A* & *B*. Table 1 illustrates these and other rules in a 'natural deduction' system.

A proof in such a system starts with a set of premises, and uses the rules to derive the conclusion. Consider the inference:

If the printer works then it has ink in its cartridges.
It does not have ink in its cartridges.
Therefore, the printer does not work.

Its proof starts with the premises expressed in the language of the sentential calculus:

1. p → i
2. ¬ i

It then proceeds as follows using the rules summarized in Table 1:

3. Suppose p (a supposition can be introduced at any point)
4. ∴ i (rule for eliminating →, lines 1 and 3)
5. ∴ i & not i (introduction of &, lines 4 and 2)
6. ∴ ¬ p (reductio ad absurdum, lines 3 and 5)

The final step is based on a rule known as reductio ad absurdum, which stipulates that if a supposition leads to a contradiction (as in line 5), then one can deny the supposition.

The third component of a logic is its semantics. Logicians assume that the truth or falsity of any sentence in the sentential calculus depends on the truth or falsity of its atomic propositions, i.e., those propositions that contain neither negation nor sentential connectives. The meaning of negation is simple: if a sentence *A* is true then ¬ *A* is false, and if *A* is false then ¬ *A* is true. Likewise, the meaning of conjunction is simple: if *A* is true and *B* is true, then *A* & *B* is true; otherwise, it is false. Logicians often lay out the meaning of a connective in a truth table, e.g.:

A	*B*	*A* and *B*
True	True	True
True	False	False
False	True	False
False	False	False

Each row in the table shows a possible combination of truth values for the sentence *A* and for the sentence *B*, and the resulting truth value of the conjunction, *A* & *B*. The first row in the table, for instance, represents the case where *A* is true and *B* is true, and so the conjunction is true.

The meaning of disjunction is likewise obvious: *A v B* is true provided that at least one of its two sentences is true, and false otherwise. The meaning of ' →' is defined in this way: *A* → *B* is true in any case except the one in which *A* is true and *B* is false. It accordingly treats the conditional about the printer as though it meant: if the printer works then it has ink in its cartridges, and if it does not work then either there is ink in its cartridges or there is not. In sum, the semantics of the sentential calculus is truth functional, i.e., the meaning of each logical term is a function that takes truth values, true or false, as its input and that outputs a single truth value.

The set of formal rules constitute a system for proving that conclusions can be derived from premises, and they are sensitive to the logical form of the premises, which is specified by the grammar. The formal rules of inference are rules for writing new patterns of symbols given certain other patterns of symbols, and the rules are sensitive to the form of the symbols, not to their meaning. A formal system accordingly operates like a computer program. When a computer program plays a game of chess, for example, the computer itself has no idea of what chess is or of what it is doing. It merely slavishly follows its program, operating on 'bits', which are symbols made up from patterns of electricity, and from time to time displaying symbols that the human users of the program can interpret as moves on a chess board. There is an intimate relation between computer programs and proofs, and programs have been written to prove theorems. The programming language, PROLOG, is itself closely related to a logical calculus.

Formal rules make no reference to validity, which is the province of the semantics: a valid inference is one that has a true conclusion given that its premises are true. We can check that an inference is valid using a truth table to consider all the possible assignments of truth values to the atomic propositions in the premises in order to show that when the premises are true, the conclusion is true too. Logicians have proved that a natural deduction system for the sentential calculus has the desirable property that any inference that can be proved using the formal rules is valid according to the semantics (the calculus is 'sound'), and the further desirable property that any inference that is valid according to the semantics is provable using the formal rules (the calculus is 'complete'). There is also a decision procedure that for any inference determines in a finite number of steps whether or not it is valid. The greatest logical discovery of the twentieth century, however, was Gödel's proof of the *incompleteness* of any logic powerful enough to express arithmetic, i.e., there are truths of arithmetic that cannot be proved using any consistent system of formal rules (see Ref 2, Chapter 7). Another major logical discovery concerned the 'first-order predicate calculus', a logic that combines the sentential calculus with rules for reasoning about the properties of individuals and relations among them. No formal system for this calculus can yield a decision procedure about the status of an inference. If an inference is valid, a proof for it can always be found in an exhaustive exploration of the possibilities. But, if an inference is invalid, no guarantee exists that a demonstration of its invalidity can be found.

ARTIFICIAL INTELLIGENCE AND PROGRAMS FOR PROOF

The practicalities of computer programming call for a decision about validity within a reasonable amount of time. The sentential calculus, however, is computationally intractable, that is, as the length of the premises increases, so the time it takes to discover a proof increases in an exponential way. Yet, given a proof, the time to check that it is correct increases only as some polynomial of the length of the premises, e.g., n^3, where n is the length of the premises.[5] So, if one could guess a proof, then the time to check it would take only a polynomial of n. The sentential calculus is accordingly NP-complete (N for 'nondeterministic', which is jargon for 'guessing', P for 'polynomial', and 'complete' because if someone discovered a deterministic way to find sentential proofs in polynomial time, a whole set of other problems would also have polynomial time solutions.[6]

Some programs for proofs use natural deduction systems.[7] But, any sentential connective can be re-expressed as a disjunction: *A & B* is equivalent to $\neg(\neg A \ v \neg B)$, and $A \rightarrow B$ is equivalent to $\neg A \ v \ B$. To try to cut down the search for a proof, early programs used just a single rule of inference, the so-called *resolution* rule[8]:

A v B
¬ B v C.
Therefore, A v C.

Yet, there is no way round the intractability of the sentential calculus: the search at each step depends on finding the appropriate pair of sentences to which to apply the resolution rule.

An alternative approach in artificial intelligence uses, not formal rules of inference, but rules that have a specific content.[9] Such programs typically have a data base of facts, which are written in a logical notation, so, for example, the sentence 'Pat is a teacher' is stated in the data base as:

(teacher Pat)

The program allows a user to interrogate the data base, and it can add new sentences to it. If the user enquires whether Pat is a teacher, the program responds, 'yes'. And if the user types in the further information: 'Pat is a reader', the program adds this new information to the data base: (reader Pat). But, suppose one wants to establish a general principle, such as: All teachers are readers. The program deals with such a generalization using a content-specific rule, which stipulates in effect: if x is a teacher, add the fact that x is a reader to the data base. The rule springs to life when anyone is described as a teacher, and adds the further information that this individual is also a reader. The same generalization, however, can be used in a different form of rule, which says in effect if the goal is to determine whether a person is a reader, set up a subgoal to show that this individual is a teacher. Now, if a user enquires: 'Is Viv a reader?', there may be no matching assertion in the data base. But, the rule springs to life to try to discover if Viv is a teacher. And, if there is a sentence to that effect in the data base, or it can be proved, then the answer to the user's question, via the new rule, is, 'yes'.

In logic, general assertions are axioms—further premises to be taken into account in proving theorems. But, these programs represent such axioms as content-specific rules. These rules either add information to the data base when an assertion is made, or deduce it from the data base in order to answer questions. But, the rules in turn can be equipped with specific heuristics for achieving particular inferential goals. Instead of a blind uniform search, such as the resolution method, the content of a problem can directly affect the way in which a search for a proof is carried out.

Computer scientists have developed another approach based on knowledge. It tackles a reasoning problem by finding a method in its data base that has been successfully applied to a similar case in the past.[10,11] According to this idea of 'case-based' reasoning, human reasoning has little to do with logic. What happens is that one inference calls to mind another. But, when an activity has been repeated often enough, it begins to function like a content-specific rule.

Knowledge-based systems offer no immediate explanation of the ability to reason about the unknown. Even if you know nothing about nondeterminism, you can make the following deduction:

If the process is nondeterministic then its runs in polynomial time.

The process is nondeterministic.

∴ It runs in polynomial time.

Theories that postulate mechanisms for dealing with specific domains[12,13] can therefore tell only part of the story of reasoning. More abstract deductive competence is necessary for logic and mathematics, and knowledge-based theories cannot immediately explain it.

As programmers developed systems for reasoning, they made several discoveries that had eluded psychologists. One discovery was that simple tacit inferences play a crucial part in the comprehension of everyday discourse. In an assertion, such as: 'If he put the parcel on the table, then it rolled off', an inference is necessary to determine that the pronoun 'it' refers to the parcel, and not to the entity most recently referred, to the table.[14] Part of the reason that no existing program can understand English in full is the lack of a sufficiently rich inferential system. A particular difficulty is that these inferences are often invalid, e.g., in the preceding case it is possible that the table itself rolled off.

Another important discovery illuminated a major difference between logic and everyday reasoning. Much human knowledge is in the form of idealizations, e.g., birds fly, dogs bark, tigers have stripes. If you learn that Tweety is a bird, then this knowledge allows you to infer:

Tweety flies.

But, if you find out that, say, Tweety is an emu or has his feet encased in concrete, you withdraw the conclusion. In contrast, logic is *monotonic*, that is to say, the addition of a new premise never entails that a prior valid conclusion should be withdrawn. In logic, one never has to say 'sorry' about a valid inference. Hence, given the following premises:

Tweety is a bird.

All birds fly.

Tweety is an emu.

No emus fly.

logic allows you to infer from the first two premises that Tweety flies, and to infer from the last two premises that Tweety does not fly. You have inferred a contradiction. But, logic does not prohibit inferences to contradictions—we already saw that a reductio ad absurdum depends on inferring a contradiction. But, the contradiction about Tweety did not derive from a supposition, and so, as far as logic is concerned, there is nothing to withdraw.

Unlike logic, everyday reasoning is not monotonic. A new premise that conflicts with a conclusion may lead you to withdraw the conclusion, and in turn to reject one of the premises from which it derives But, which premise? Logic cannot tell you. So, which premise should you withdraw in the Tweety example? The answer is clear: it is not true that all birds fly. What you should like to maintain, however, is that birds typically fly. English expresses this proposition in a generic assertion that omits the quantifier 'all' or its equivalents: Birds fly. Generics are useful, because they tolerate exceptions, and yet they are true for typical cases.

Artificial intelligencers accordingly sought some way of formulating *nonmonotonic* reasoning systems in which the rebuttal of a conclusion did not call for the withdrawal of useful generic premises, such as 'birds fly'. Researchers tried out various ideas, and there developed a cottage industry of nonmonotonic reasoning systems.[15] A psychologically plausible idea is due to Minsky.[16] He formulated the notion of a 'default value', e.g., a bird flies by default. In other words, you can infer that a bird flies, unless there is evidence to the contrary. In which case, you withdraw the conclusion, but not the generic idealization. This notion was embodied in 'object-oriented' programming languages, which allow programmers to set up hierarchies of concepts, so that birds include emus; birds fly, but this default is overruled in the case of emus, which do not fly. These systems allow you to have your cake and eat it. And human reasoning often appears to rely on defaults, which are arguably the correct semantic analysis for generic assertions. Hence, a statement such as:

> Smoking causes cancer

allows for exceptions. Because the assertion allows exceptions, some theories postulate that causation is itself a probabilistic notion (e.g., Ref 17). But, it is the generic nature of the assertion that gives rise to the tolerance of exceptions. A claim such as:

> All cases of smoking cause cancer

is not probabilistic, and so the concept of cause is not probabilistic, either.[18]

One final point about nonmonotonicity is that it also occurs in cases where beliefs do have to be revised. Suppose you believe the following propositions:

> If someone pulled the trigger then the pistol fired.
>
> Someone did pull the trigger.

And then you learn that the pistol did not fire. You are likely to try to figure out an explanation of the sequence of events. You might suppose that a prudent person emptied the bullets from the pistol. In this case, your explanation overturns the first of your two premises.[19]

PSYCHOLOGICAL THEORIES OF DEDUCTION

The task for psychologists once seemed to be to identify the particular logic that people have in their heads—an idea going back to the ancient doctrine that the laws of logic are the laws of thought. The problem was the vast number of different logics, and the variety of different ways of formalizing them. Logicians have proved that there are infinitely many distinct modal logics, which deal with possibility and necessity. Nevertheless, theorists argued for a century that logic is a theory of deductive competence.[20] Inhelder and Piaget (Ref 21, p. 305) went so far as to claim that reasoning is nothing more than the sentential calculus itself. Others similarly argued that deductive performance depends on formal rules of inference akin to those in Table 1 (e.g., Refs 22 and 23). This view has adherents in psychology (e.g., Refs 24 and 25), in linguistics (e.g., Ref 26), in philosophy (e.g., Ref 27), and in artificial intelligence (e.g., Ref 28). But, three major theoretical difficulties confront any psychological theory based on logic.

The first difficulty is that logic allows an infinite number of different conclusions to follow validly from any premises. Consider, for instance, the following premises:

> If the printer works then it has ink in its cartridges.
>
> The printer works.

They validly imply the infinite list of conclusions beginning:

> The printer has ink in its cartridges.
>
> The printer has ink in its cartridges and it has ink in its cartridges.
>
> The printer has ink in its cartridges and it has ink in its cartridges and it has ink in its cartridges.

Of course the last two conclusions are preposterous. No sane individual—other than a logician—is likely to draw them. Yet they are valid deductions. Hence, logic alone cannot be a theory of deductive reasoning (pace Ref 21). Likewise, in logic, the response that nothing follows from the premises is always wrong. Yet, people are sensible. They do not draw just any valid conclusion, and for some premises the sensible response is that nothing follows (of any interest). Logic is at best an incomplete theory of deductive reasoning, because it has nothing to say about which logical conclusions are sensible. What naïve individuals—those who have not mastered logic—tend to infer are conclusions that do not throw information away by adding disjunctive alternatives to those the premises support, that simplify matters rather than add redundant propositions, and that make explicit what was only implicit in the premises (Ref 29, p. 22). This account of deductive competence has yet to be overturned. And none of its principles can be derived from logic alone.

The second difficulty for logic-based theories is the gap between formal logic and natural language. Logic concerns implications between *sentences*. Everyday reasoning concerns implications between the *propositions* that sentences express, and most sentences can express many different propositions. Consider a sentence such as:

If she played a game then she did not play *that*.

The sentence expresses different propositions depending on whom 'she' refers to, and on the game to which the speaker points on uttering 'that'. The grammar of a natural language captures the grammatical form of the sentences in the language, but grammatical form is not necessarily equivalent to logical form. Consider all the possible inferences of the following grammatical form:

If *A* then not *B*.
B.
Therefore, not *A*.

If she is in Brazil then she is not in Oslo.
She is in Oslo.
Therefore, she is not in Brazil.

On the other hand, you are likely to balk at this case:

If she is in Brazil then she is not in Rio.
She is in Rio.
Therefore, she is not in Brazil.

You may say that there is a missing premise: if she is in Rio then she is in Brazil. Alas, all this premise does is to make the three premises self-contradictory, and, as we have already seen, in logic a contradiction does not force us to withdraw a conclusion. An index of the difficulty of determining the logical form of propositions expressed in natural language is that no algorithm exists for automatically doing the job. The one program implementing a psychological theory based on formal rules accordingly calls for users themselves to provide the logical form of premises and conclusions.[24]

Arguments in books or articles do not resemble formal proofs. It is almost impossible to determine their logical form, and hence to use logic to assess their validity. This problem has led some theorists to argue that logic is irrelevant to the inferences of daily life[30] (see also the ILACT movement for 'Informal Logic And Critical Thinking'). When logicians do analyze everyday inferences, they typically discover that their logical forms are ambiguous.[31] Yet, validity is important in daily life, and, as we will see, it can be assessed without having to determine the logical form of arguments.

The third difficulty for logic-based theories is that the validity of many inferences in daily life depends on the meanings of words other than logical terms. Consider the following deduction:

Alan is taller than Betty.
Betty is taller than Charlie.
Therefore, Alan is taller than Charlie.

On the definition of validity at the start of this article, the inference is valid, i.e., if its premises are true then so too is its conclusion. But, the inference cannot be proved in logic without the addition of a missing premise to the effect that: for any individuals, x, y, z, if x is taller than y, and y is taller than z, then x is taller than z. This premise amounts to an axiom, or *meaning postulate*, because it captures the transitivity of the relation, 'is taller than'. The task of specifying the required set of meaning postulates is formidable, and has yet to be done. One reason is that a simple change in tense in one of the two premises of the preceding example suffices to invalidate transitivity:

Alan is taller than Betty.
Betty *was* taller than Charlie.

The past tense alerts you to the possibility that Charlie has grown, and may now be taller than Alan. Indeed, naïve individuals balk at drawing the transitive conclusion in such cases.[32] But, matters are still worse in the case of certain spatial inferences.[33] If three individuals, Matthew, Mark, and Luke, are seated along one side of a rectangular table, then the fact that Matthew is to Mark's right, and Mark is to Luke's right, suffices to infer validly that Matthew is to Luke's right. But, if they are equally spaced around a small circular table, the inference is no longer valid, because Matthew is opposite Luke. Depending on the radius of the table and the closeness of the seating, transitivity may hold over, say, four individuals but break down over five individuals, hold over five individuals but break down over six, and so on, up to an indefinite number of individuals. Each of these different cases calls for its own meaning postulates, and so one wonders whether there might not be a better way for handing the deductions. There is; and we now turn to it.

The theory of mental models postulates that deductive reasoning depends not on formal rules, but on the use of meaning, reference, and knowledge to construct mental models of the possibilities in which premises hold.[29,34,35] Mental models have three essential characteristics:

1. Each model represents a possibility, just as each true row in a truth table corresponds to a possibility. Hence, the inclusive disjunction:

   There is a circle or there is not a triangle.

   calls for three mental models, shown here on separate lines, which represent the three possibilities:

   $$\begin{array}{ll} \text{o} & \\ & \neg\,\Delta \\ \text{o} & \neg\,\Delta \end{array}$$

2. The principle of truth: to minimize the load on working memory, mental models represent only what is true, and not what is false. This principle applies at two levels. First, each mental model represents a true possibility. Hence, the set of models above does not represent the case in which the disjunction as a whole is false. Second, atomic propositions and the negations of atomic propositions in premises are represented in a mental model only when they are true in the corresponding possibility. Hence, the first model in the set above represents the possibility that there is a circle, but it does not make explicit that in this case it is false that there is not a triangle, i.e., there *is* a triangle. When individuals reason intuitively they rely on mental models, typically just a single model.[33] Only in reasoning in a more deliberate way do they use fully explicit models, which represent both what is true and what is false.
3. The parts of a model correspond to the parts of what it represents, and the structure of the model corresponds to the structure of what it represents. Thus, a mental model is like an architect's model of a building. It can also be used in some cases to construct a visual image, though many mental models are not visualizable.

The theory postulates that models are based on descriptions, on perception, and on knowledge, including, for example, knowledge of the size and seating arrangements in the earlier case of the round table. Reasoners formulate a conclusion that holds in the models and that was not explicitly asserted in any single premise. The strength of the inference depends on the proportion of the models in which the conclusion holds. A conclusion that holds in all the models is necessary given the premises.[29] A conclusion that holds in at least one model is possible given the premises.[36] And a conclusion that holds in most of the models is probable given the premises.[37] Models accordingly provide a unified theory of logical, modal, and probabilistic reasoning—at least the sort of probabilistic reasoning that depends on adding the probabilities of the different ways in which an event can occur.

Do logically untrained individuals reason using formal rules, content-specific rules, mental models, or some other system? Experiments have established crucial phenomena that may help readers to decide. The model theory predicts that reasoners should be able to reason from exclusive disjunctions such as: *A or else B, but not both*, more easily than from inclusive disjunctions, such as: *A or B, or both*. Exclusive disjunctions hold in only two possibilities, whereas inclusive disjunctions hold in three possibilities. In contrast, formal rule theories have rules only for inclusive disjunctions, and so proofs from exclusive disjunctions call for extra steps in which the exclusive disjunction is paraphrased as: (*AvB*) & ¬ (*A* & *B*), and then one of these clauses is detached as an intermediate conclusion. The evidence in several sorts of study shows that deductive reasoning from exclusive disjunctions is easier than from inclusive disjunctions (for a review, see Ref 29). Formal or content-specific rule theories could, of course, introduce a rule for exclusive disjunctions, but they have no way to *predict* that the use of this rule should be easier than the use of the rule for inclusive disjunctions.

Another phenomenon concerns the use of counterexamples. The model theory predicts that individuals can refute invalid inferences by envisaging counterexamples to their conclusions. A counterexample is a possibility in which the premises hold, but the conclusion does not. In contrast, current formal rule theories (e.g., Refs 24 and 25) make no use of counterexamples. But, when participants had to write down justifications for their evaluations of inferences, they used counterexamples, especially to refute conclusions that were consistent with the premises but that did not follow from them.[38] Studies of other sorts of reasoning have shown that individuals spontaneously use counterexamples.[39] And the inferences in these studies cannot be based on general knowledge.

Another phenomenon arose from an apparent bug in a computer program implementing the model theory. In fact, the bug was in the author's mind rather than the program. What it showed was that the principle of truth predicted that certain inferences should elicit systematic fallacies. These fallacies occur in many domains of reasoning (e.g., Refs 34 and 40). A simple case due to Khemlani[41] concerns a rule at a restaurant:

Only one of these assertions is true:

1. You can have the bread.
2. You can have the soup or else the salad, but not both.

Let us say that you decide to have the bread. Is it possible to have the soup and the salad as well?

If you answered 'no', then you are like the participants in the experiments.[41] Yet, the inference is illusory. The principle of truth predicts that you form these mental models of what are you allowed according to the restaurant's rule:

bread
soup
salad

But, the fully explicit models of the rule, which represent both what is true and what is false, show that if you have the bread according to part (1) of the rule then it is possible for you to have both the soup and the salad, because taking them both would be a case in which part (2) of the rule would be false. The majority of participants in the studies succumbed to the fallacies. They are so compelling that they are known as 'illusory inferences'. People go wrong because they cannot cope with what is false.[42] If theories that postulate only the sorts of rule in Table 1 are correct, then the illusions should not occur. But, they do occur, and hence jeopardize these theories. Valid principles cannot explain systematic invalidity. However, Rips[24] has suggested that individuals may have erroneous rules of inference. The problem with this proposal is to formulate a theory that predicts the systematic pattern of real and illusory inferences. Once again, theories based on content-specific rules or knowledge fail to predict the phenomena.

CONCLUSION

Deductive reasoning is critical to many human activities. This paper, for example, has relied on at least one valid inference of the form:

> If theory *X* is correct then phenomenon *Y* should not occur.
>
> But, phenomenon *Y* does occur.
>
> Therefore, theory *X* is not correct.

Despite the ubiquity of such inferences, some psychologists have argued that logical validity is the wrong criterion for human reasoning, and that it should be replaced by probability.[43] The probabilistic approach yields excellent accounts of some experiments—though not those summarized in the previous section. And, contrary to its claims, individuals *are* sensitive to the difference between valid deductions and probabilistic conclusions.[44] Other current theories of reasoning postulate, as we saw earlier, that it relies on content-specific rules, or even innate mental modules for specific topics, such as checking whether someone is cheating you[13]. The difficulty in assessing these accounts is that theorists have yet to specify the complete set of modules, their detailed workings, or the principles by which problems trigger a particular module. What remains highly controversial are the meaning and mental representation of *if-then* assertions in daily life,[3,45–47] estimates of their probability,[48–50] and deductive reasoning from them.[51–54] Theoretical generality and the evidence lean toward the model theory.[55–57] But, of course, a new theory could lead to the discovery of robust phenomena contrary to the model theory. The theory postulates that counterexamples can overturn theories, and so it will at least account for its own demise.

REFERENCES

1. Stanovich KE. *Who is Rational? Studies of Individual Differences in Reasoning*. Mahwah, NJ: Erlbaum; 1999.
2. Jeffrey R. *Formal Logic: Its Scope and Limits*. 2nd ed. New York: McGraw-Hill; 1981.
3. Johnson-Laird PN, Byrne RMJ. Conditionals: a theory of meaning, pragmatics, inference. *Psychol Rev* 2002, 109:646–678.
4. Gentzen G. Investigations into logical deduction. In: Szabo ME, ed. and trans. *The Collected Papers of Gerhard Gentzen*. Amsterdam: North-Holland; 1969, 68–131.
5. Cook SA. The complexity of theorem proving procedures. In: *Proceedings 3rd Annual Association for Computing Machinery Symposium on Theory of Computing*; 1971, 151–158.
6. Garey M, Johnson D. *Computers and Intractability: A Guide to the Theory of NP-completeness*. San Francisco, CA: Freeman; 1979.
7. Russell SJ, Norvig P. *Artificial Intelligence: A Modern Approach*. 2nd ed. Upper Saddle River, NJ: Pearson Education; 2003.
8. Robinson JA. *Logic: Form and Function* Edinburgh: Edinburgh University Press; 1979.
9. Hewitt C. PLANNER: A Language for Proving Theorems in Robots. *Int Joint Conf Artif Intell* 1969, 2:295–302.
10. Riesbeck CK, Schank RC. *Inside Case-Based Reasoning*. Hillsdale, NJ: Erlbaum; 1989.
11. Kolodner J. *Case-based Reasoning*. San Mateo, CA: Morgan Kaufman; 1993.
12. Cheng PW, Holyoak KJ. Pragmatic reasoning schemas. *Cognit Psychol* 1985, 17:391–416.
13. Cosmides L. The logic of social exchange: has natural selection shaped how humans reason? *Cognition* 1989, 31:187–276.
14. Charniak E. Jack and Janet in search of a theory of knowledge. *Third Int Joint Conf Artif Intell* 1973, 331–337.
15. Brewka G, Dix J, Konolige K. *Nonmonotonic Reasoning: An Overview*. Stanford, CA: CLSI Publications, Stanford University; 1997.
16. Minsky M. Frame-system theory. In: Johnson-Laird PN, Wason PC, eds. *Thinking: Readings in Cognitive Science*. Cambridge: Cambridge University Press; 1977, 355–376.
17. Cheng PW. From covariation to causation: a causal power theory. *Psychol Rev* 1997, 104:367–405.
18. Goldvarg Y, Johnson-Laird PN. Naïve causality: a mental model theory of causal meaning and reasoning. *Cogn Sci* 2001, 25:565–610.
19. Johnson-Laird PN, Girotto V, Legrenzi P. Reasoning from inconsistency to consistency. *Psychol Rev* 2004, 111:640–661.
20. Boole G. *An Investigation of the Laws of Thought on Which are Founded the Mathematical Theories of Logic and Probabilities*. London: Macmillan; 1854.
21. Inhelder B, Piaget J. *The Growth of Logical Thinking from Childhood to Adolescence*. London: Routledge & Kegan Paul; 1958.
22. Osherson DN. *Logical Abilities in Children*, Vols 1–4. Hillsdale, NJ: Erlbaum; 1974–1976.
23. Braine MDS. On the relation between the natural logic of reasoning and standard logic. *Psychol Rev* 1978, 85:1–21.
24. Rips LJ. *The Psychology of Proof*. Cambridge, MA: MIT Press; 1994.
25. Braine MDS, O'Brien DP, eds. *Mental Logic*. Mahwah, NJ: Erlbaum; 1998.
26. Sperber D, Wilson D. *Relevance: Communication and Cognition*. Oxford: Basil Blackwell; 1986.
27. Pollock J. *How to Build a Person: A Prolegomenon* Cambridge, MA: MIT Bradford Books; 1989.
28. Wos L. *Automated Reasoning: 33 Basic Research Problems*. Englewood Cliffs, NJ: Prentice-Hall; 1988.
29. Johnson-Laird PN, Byrne RMJ. *Deduction*. Hillsdale, NJ: Erlbaum; 1991.
30. Toulmin SE. *The Uses of Argument* Cambridge: Cambridge University Press; 1958.
31. Keene GB. *The Foundations of Rational Argument*. Lampeter, Wales: Edwin Mellen Press; 1992.
32. Goodwin GP, Johnson-Laird PN. Transitive and pseudo-transitive inferences. *Cognition* 2008, 108:320–352.
33. Johnson-Laird PN. *Mental Models: Towards a Cognitive Science of Language, Inference and Consciousness*. Cambridge, MA: Cambridge University Press, Harvard University Press; 1983.
34. Johnson-Laird PN. *How We Reason* New York: Oxford University Press; 2006.
35. Polk TA, Newell A. Deduction as verbal reasoning. *Psychol Rev* 1995, 102:533–566.
36. Bell V, Johnson-Laird PN. A model theory of modal reasoning. *Cogn Sci* 1998, 22:25–51.
37. Johnson-Laird PN, Legrenzi P, Girotto V, Legrenzi M, Caverni J-P. Naïve probability: a mental model theory of extensional reasoning. *Psychol Rev* 1999, 106:62–88.
38. Johnson-Laird PN, Hasson U. Counterexamples in sentential reasoning. *Mem Cognit* 2003, 31:1105–1113.
39. Bucciarelli M, Johnson-Laird PN. Strategies in syllogistic reasoning. *Cogn Sci* 1999, 23:247–303.

40. Johnson-Laird PN, Savary F. Illusory inferences: a novel class of erroneous deductions. *Cognition* 1999, 71:191–229.

41. Khemlani S, Johnson-Laird PN. Disjunctive illusory inferences and how to eliminate them. 2009., Under submission.

42. Yang Y, Johnson-Laird PN. How to eliminate illusions in quantified reasoning. *Mem Cognit* 2000, 28:1050–1059.

43. Oaksford M, Chater N. *Bayesian Rationality: The Probabilistic Approach to Human Reasoning*. New York: Oxford University Press; 2007.

44. Markovits H, Handley SJ. Is inferential reasoning just probabilistic reasoning in disguise? *Mem Cognit* 2005, 33:1315–1323.

45. Evans JStBT, Over DE. *If. 2004*. New York: Oxford University Press; 2004.

46. Byrne RMJ. *The Rational Imagination: How People Create Alternatives to Reality*. Cambridge, MA: MIT Press; 2005.

47. Barrouillet P, Gauffroy C, Lecas J-F. Mental models and the suppositional account of conditionals. *Physiol Rev* 2008, 115:760–772.

48. Evans JStBT, Handley SH, Over DE. Conditionals and conditional probability. *J Exp Psychol Learn Mem Cognit* 2003, 29:321–335.

49. Girotto V, Johnson-Laird PN. The probability of conditionals. *Psychologia* 2004, 47:207–225.

50. Over DE, Hadjchristidis C, Evans JStBT, Handley SE, Sloman S. The probability of causal conditionals. *Cognit Psychol* 2007, 54:62–97.

51. García-Madruga JA, Moreno S, Carriedo N, Gutiérrez F, Johnson-Laird PN. Are conjunctive inferences easier than disjunctive inferences? A comparison of rules and models. *Q J Exp Psychol* 2001, 54A:613–632.

52. Santamaria C, Espino O. Conditionals and directionality: on the meaning of if versus only if. *Q J Exp Psychol* 2002, 55A:41–57.

53. Quelhas AC, Byrne RMJ. Reasoning with deontic and counterfactual conditionals. *Think Reason* 2003, 9:43–66.

54. Quelhas AC, Johnson-Laird PN, Juhos C. The modulation of conditional assertions and its effects on reasoning. *Q J Exp Psychol* 2010. In press.

55. Schroyens W, Schaeken W. A critique of Oaksford, Chater and Larkin's (2000) conditional probability model of conditional reasoning. *J Exp Psychol Learn Mem Cognit* 2003, 29:140–149.

56. Schroyens W, Schaeken W, d'Ydewalle G. The processing of negations in conditional reasoning: A meta-analytic case study in mental model and/or mental logic theory. *Think Reason* 2001, 7:121–172.

57. Oberauer K. Reasoning with conditionals: a test of formal models of four theories. *Cognit Psychol* 2006, 53:238–283.