Deduction

P.N. Johnson-Laird
Department of Psychology
Princeton University
Green Hall
Princeton
NJ 08544, USA.

Ruth M.J. Byrne
Department of Psychology
Trinity College
University of Dublin
Dublin 2, Ireland

LAWRENCE ERLBAUM ASSOCIATES, PUBLISHERS
Hove (UK) Hillsdale (USA)

CHAPTER 2

The Cognitive Science of Deduction

The late Lord Adrian, the distinguished physiologist, once remarked that if you want to understand *how* the mind works then you had better first ask *what* it is doing. This distinction has become familiar in cognitive science as one that Marr (1982) drew between a theory at the "computational level" and a theory at the "algorithmic level". A theory at the computational level characterizes what is being computed, why it is being computed, and what constraints may assist the process. Such a theory, to borrow from Chomsky (1965), is an account of human competence. And, as he emphasizes, it should also explain how that competence is acquired. A theory at the algorithmic level specifies how the computation is carried out, and ideally it should be precise enough for a computer program to simulate the process. The algorithmic theory, to borrow again from Chomsky, should explain the characteristics of human performance—where it breaks down and leads to error, where it runs smoothly, and how it is integrated with other mental abilities.

We have two goals in this chapter. Our first goal is to characterize deduction at the computational level. Marr criticized researchers for trying to erect theories about mental processes without having stopped to think about what the processes were supposed to compute. The same criticism can be levelled against many accounts of deduction, and so we shall take pains to think about its function: what the mind computes, what purpose is served, and what constraints there are on the process. Our second goal is to examine existing algorithmic theories. Here, experts in several domains of enquiry have something to say. Linguists have considered the logical form of sentences in natural language.

Computer scientists have devised programs that make deductions, and, like philosophers, they have confronted discrepancies between everyday inference and formal logic. Psychologists have proposed algorithmic theories based on their experimental investigations. We will review work from these disciplines in order to establish a preliminary account of deduction—to show what it is, and to outline theories of how it might be carried out by the mind.

DEDUCTION: A THEORY AT THE COMPUTATIONAL LEVEL

What happens when people make a deduction? The short answer is that they start with some information—perceptual observations, memories, statements, beliefs, or imagined states of affairs—and produce a novel conclusion that follows from them. Typically, they argue from some initial propositions to a single conclusion, though sometimes merely from one proposition to another. In many practical inferences, their starting point is a perceived state of affairs and their conclusion is a course of action. Their aim is to arrive at a valid conclusion, which is bound to be true given that their starting point is true.

One long-standing controversy concerns the extent to which people are logical. Some say that logical error is impossible: deduction depends on a set of universal principles applying to any content, and everyone exercises these principles infallibly. This idea seems so contrary to common sense that, as you might suspect, it has been advocated by philosophers (and psychologists). What seems to be an invalid inference is nothing more than a valid inference from other premises (see Spinoza, 1677; Kant, 1800). In recent years, Henle (1962) has defended a similar view. Mistakes in reasoning, she claims, occur because people forget the premises, re-interpret them, or import extraneous material. "I have never found errors," she asserts, "which could unambiguously be attributed to faulty reasoning" (Henle, 1978). In all such cases, the philosopher L. J. Cohen (1981) has concurred, there is some malfunction of an information-processing mechanism. The underlying competence cannot be at fault. This doctrine leads naturally to the view that the mind is furnished with an inborn logic (Leibniz, 1765; Boole, 1854). These authors, impressed by the human invention of logic and mathematics, argue that people must think rationally. The laws of thought are the laws of logic.

Psychologism is a related nineteenth century view. John Stuart Mill (1843) believed that logic is a generalization of those inferences that people judge to be valid. Frege (1884) attacked this idea: logic may

ultimately depend on the human mind for its discovery, but it is not a subjective matter; it concerns objective relations between propositions.

Other commentators take a much darker view about logical competence. Indeed, when one contemplates the follies and foibles of humanity, it seems hard to disagree with Dostoyevsky, Nietzsche, Freud, and those who have stressed the irrationality of the human mind. Yet this view is reconcilable with logical competence. Human beings may desire the impossible, or behave in ways that do not optimally serve their best interests. It does not follow that they are incapable of rational thought, but merely that their behaviour is not invariably guided by it.

Some psychologists have proposed theories of reasoning that render people inherently irrational (e.g. Erickson, 1974; Revlis, 1975; Evans, 1977a). They may draw a valid conclusion, but their thinking is not properly rational because it never makes a full examination of the consequences of premises. The authors of these theories, however, provide no separate account of deduction at the computational level, and so they might repudiate any attempt to ally them with Dostoyevsky, Nietzsche, and Freud.

Our view of logical competence is that people are rational in principle, but fallible in practice. They are able to make valid deductions, and moreover they sometimes *know* that they have made a valid deduction. They also make invalid deductions in certain circumstances. Of course, theorists can explain away these errors as a result of misunderstanding the premises or forgetting them. The problem with this manoeuvre is that it can be pushed to the point where no possible observation could refute it. People not only make logical mistakes, they are even prepared to concede that they have done so (see e.g. Wason and Johnson-Laird, 1972; Evans, 1982). These meta-logical intuitions are important because they prepare the way for the invention of self-conscious methods for checking validity. Thus, the development of logic as an intellectual discipline requires logicians to be capable of sound pre-theoretical intuitions. Yet, logic would hardly have been invented if there were never occasions where people were uncertain about the status of an inference. Individuals do sometimes formulate their own principles of reasoning, and they also refer to deductions in a meta-logical way. They say, for example: "It seems to follow that Arthur is in Edinburgh, but he isn't, and so I must have argued wrongly." These phenomena merit study like other forms of meta-cognition (see e.g. Flavell, 1979; Brown, 1987). Once the meta-cognitive step is made, it becomes possible to reason at the meta-meta-level, and so on to an arbitrary degree. Thus, cognitive psychologists and devotees of logical puzzles (e.g. Smullyan, 1978; Dewdney, 1989) can in turn make inferences about meta-cognition. A

psychological theory of deduction therefore needs to accommodate deductive competence, errors in performance, and meta-logical intuitions (cf. Simon, 1982; Johnson-Laird, 1983; Rips, 1989).

Several ways exist to characterize deductive competence at the computational level. Many theorists—from Boole (1847) to Macnamara (1986)—have supposed that logic itself is the best medium. Others, however, have argued that logic and thought differ. Logic is *monotonic*, i.e. if a conclusion follows from some premises, then no subsequent premise can invalidate it. Further premises lead monotonically to further conclusions, and nothing ever subtracts from them. Thought in daily life appears not to have this property. Given the premises:

Alicia has a bacterial infection.
If a patient has a bacterial infection, then the preferred treatment for the patient is penicillin.

it follows validly:

Therefore, the preferred treatment for Alicia is penicillin.

But, if it is the case that:

Alicia is allergic to penicillin.

then common-sense dictates that the conclusion should be withdrawn. But it still follows validly in logic. This problem suggests that some inferences in daily life are "non-monotonic" rather than logically valid, i.e. their conclusions can be withdrawn in the light of subsequent information. There have even been attempts to develop *formal* systems of reasoning that are non-monotonic (see e.g. McDermott and Doyle, 1980). We will show later in the book that they are unnecessary. Nevertheless, logic cannot tell the whole story about deductive competence.

A theory at the computational level must specify what is computed, and so it must account for what deductions people actually make. Any set of premises yields an infinite number of valid conclusions. Most of them are banal. Given the premises:

Ann is clever.
Snow is white.

the following conclusions are all valid:

Ann is clever and snow is white.
Snow is white and Ann is clever and snow is white.

They must be true given that the premises are true. Yet no sane individual, apart from a logician, would dream of drawing them. Hence, when reasoners make a deduction in daily life, they must be guided by more than logic. The evidence suggests that at least three extra-logical constraints govern their conclusions.

The first constraint is *not* to throw semantic information away. The concept of semantic information, which can be traced back to medieval philosophy, depends on the proportion of possible states of affairs that an assertion rules out as false (see Bar-Hillel and Carnap, 1964; Johnson-Laird, 1983). Thus, a conjunction, such as:

Joe is at home and Mary is at her office.

conveys more semantic information (i.e. rules out more states of affairs) than only one of its constituents:

Joe is at home.

which, in turn, conveys more semantic information than the inclusive disjunction:

Joe is at home or Mary is at her office, or both.

A valid deduction cannot increase semantic information, but it can decrease it. One datum in support of the constraint is that valid deductions that do decrease semantic information, such as:

Joe is at home.
Therefore, Joe is at home or Mary is at her office, or both.

seem odd or even improper (see Rips, 1983).

A second constraint is that conclusions should be more parsimonious than premises. The following argument violates this constraint:

Ann is clever.
Snow is white.
Therefore, Ann is clever and snow is white.

In fact, logically untutored individuals declare that there is no valid

conclusion from these premises. A special case of parsimony is not to draw a conclusion that asserts something that has just been asserted. Hence, given the premises:

If James is at school then Agnes is at work.
James is at school.

the conclusion:

James is at school and Agnes is at work.

is valid, but violates this principle, because it repeats the categorical premise. This information can be taken for granted and, as Grice (1975) argued, there is no need to state the obvious. The development of procedures for drawing parsimonious conclusions is a challenging technical problem in logic. We present a solution to it, which is based on our psychological theory, in Chapter 9.

A third constraint is that a conclusion should, if possible, assert something new, i.e., something that was not explicitly stated in the premises. Given the premise:

Mark is over six feet tall and Karl is taller than him.

the conclusion:

Karl is taller than Mark, who is over six feet tall.

is valid but it violates this constraint because it asserts nothing new. In fact, ordinary reasoners spontaneously draw conclusions that establish relations that are not explicit in the premises.

When there is no valid conclusion that meets the three constraints, then logically naive individuals say, "nothing follows" (see e.g. Johnson-Laird and Bara, 1984). Logically speaking, the response is wrong. There are always conclusions that follow from any premises. The point is that there is no valid conclusion that meets the three constraints. We do not claim that people are aware of the constraints or that they are mentally represented in any way. They may play no direct part in the process of deduction, which for quite independent reasons yields deductions that conform to them (Johnson-Laird , 1983, Ch. 3). In summary, our theory of deductive competence posits rationality, an awareness of rationality, and a set of constraints on the conclusions that people draw for themselves. *To deduce is to maintain semantic information, to simplify, and to reach a new conclusion.*

FORMAL RULES: A THEORY AT THE ALGORITHMIC LEVEL

Three main classes of theory about the process of deduction have been proposed by cognitive scientists:

1. Formal rules of inference.
2. Content-specific rules of inference.
3. Semantic procedures that search for interpretations (or mental models) of the premises that are counterexamples to conclusions.

Formal theories have long been dominant. Theorists originally assumed without question that there is a mental logic containing formal rules of inference, such as the rule for modus ponens, which are used to derive conclusions. The first psychologist to emphasize the role of logic was the late Jean Piaget (see e.g. Piaget, 1953). He argued that children internalize their own actions and reflect on them. This process ultimately yields a set of "formal operations", which children are supposed to develop by their early teens. Inhelder and Piaget (1958, p.305) are unequivocal about the nature of formal operations. They write:

> No further operations need be introduced since these operations correspond to the calculus inherent to the algebra of propositional logic. In short, reasoning is nothing more than the propositional calculus itself.

There are grounds for rejecting this account: we have already demonstrated that deductive competence must depend on more than pure logic in order to rule out banal, though valid, conclusions. Moreover, Piaget's logic was idiosyncratic (see Parsons, 1960; Ennis, 1975; Braine and Rumain, 1983), and he failed to describe his theory in sufficient detail for it to be modelled in a computer program. He had a genius for asking the right questions and for inventing experiments to answer them, but the vagueness of his theory masked its inadequacy perhaps even from Piaget himself. The effort to understand it is so great that readers often have no energy left to detect its flaws.

Logical Form in Linguistics

A more orthodox guide to logical analysis can be found in linguistics. Many linguists have proposed analyses of the logical form of sentences, and often presupposed the existence of formal rules of inference that

enable deductions to be derived from them. Such analyses were originally inspired by transformational grammar (see e.g. Leech, 1969; Seuren, 1969; Johnson-Laird, 1970; Lakoff, 1970; Keenan, 1971; Harman, 1972; Jackendoff, 1972). What these accounts had in common is the notion that English quantifiers conform to the behaviour of logical quantifiers only indirectly. As in logic, a universal quantifier within the scope of a negation:

Not all of his films are admired.

is equivalent to an existential quantifier outside the scope of negation:

Some of his films are not admired.

But, unlike logic, natural language has no clear-cut devices for indicating scope. A sentence, such as:

Everybody is loved by somebody.

has two different interpretations depending on the relative scopes of the two quantifiers. It can mean:

Everybody is loved by somebody or other.

which we can paraphrase in "Loglish" (the language that resembles the predicate calculus) as:

For any x, there is some y, such that if x is a person then y is a person, and x is loved by y.

It can also mean:

There is somebody whom everybody is loved by.
(There is some y, for any x, such that y is a person and if x is a person, then x is loved by y.)

Often, the order of the quantifiers in a sentence corresponds to their relative scopes, but sometimes it does not, e.g.:

No-one likes some politicians.
(For some y, such that y is a politician, no x is a person and x likes y.)

where the first quantifier in the sentence is within the scope of the second.

Theories of logical form have more recently emerged within many different linguistic frameworks, including Chomsky's (1981) "government and binding" theory, Montague grammar (Cooper, 1983), and Kamp's (1981) theory of discourse representations. The Chomskyan theory postulates a separate mental representation of logical form (LF), which makes explicit such matters as the scope of the quantifiers, and which is transformationally derived from a representation of the superficial structure of the sentence (S-structure). The sentence, "Everybody is loved by somebody", has two distinct logical forms analogous to those above. The first corresponds closely to the superficial order of the quantifiers, and the second is derived by a transformation that moves the existential quantifier, "somebody", to the front—akin to the sentence:

Somebody, everybody is loved by.

This conception of logical form is motivated by linguistic considerations (see Chomsky, 1981; Hornstein, 1984; May, 1985). Its existence as a level of syntactic representation, however, is not incontrovertible. The phenomena that it accounts for might be explicable, as Chomsky has suggested (personal communication, 1989), by enriching the representation of the superficial structure of sentences.

Logical form is, of course, a necessity for any theory of deduction that depends on formal rules of inference. Kempson (1988) argues that the mind's inferential machinery is formal, and that logical form is therefore the interface between grammar and cognition. Its structures correspond to those of the deductive system, but, contrary to Chomskyan theory, she claims that it is not part of grammar, because general knowledge can play a role in determining the relations it represents. For example, the natural interpretation of the sentence:

Everyone got into a taxi and chatted to the driver.

is that each individual chatted to the driver of his or her taxi. This interpretation, however, depends on general knowledge, and so logical form is not purely a matter of grammar. Kempson links it to the psychological theory of deduction advocated by Sperber and Wilson (1986). This theory depends on formal rules of inference, and its authors have sketched some of them within the framework of a "natural deduction" system.

One linguist, Cooper (1983), treats scope as a semantic matter, i.e. within the semantic component of an analysis based on Montague grammar, which is an application of model-theoretic semantics to language in general. A different model-theoretic approach, "situation semantics", is even hostile to the whole notion of reasoning as the formal manipulation of formal representations (Barwise, 1989; Barwise and Etchemendy, 1989a,b).

Formal Logic in Artificial Intelligence

Many researchers in artificial intelligence have argued that the predicate calculus is an ideal language for representing knowledge (e.g. Hayes, 1977). A major discovery of this century, however, is that there cannot be a full decision procedure for the predicate calculus. In theory, a proof for any valid argument can always be found, but no procedure can be guaranteed to demonstate that an argument is invalid. The procedure may, in effect, become lost in the space of possible derivations. Hence, as it grinds away, there is no way of knowing if, and when, it will stop. One palliative is to try to minimize the search problem for valid deductions by reducing the number of formal rules of inference. In fact, one needs only a single rule to make any deduction, the so-called "resolution rule" (Robinson, 1965):

A or B, or both
C or not-B, or both
∴ A or C, or both.

The rule is not intuitively obvious, but consider the following example:

Mary is a linguist or Mary is a psychologist.
Mary is an experimenter or Mary is not a psychologist.
Therefore, Mary is a linguist or Mary is an experimenter.

Suppose that Mary is not a psychologist, then it follows from the first premise that she is a linguist; now, suppose that Mary is a psychologist, then it follows from the second premise that she is an experimenter. Mary must be either a psychologist or not a psychologist, and so she must be either a linguist or an experimenter.

Table 2.1 summarizes the main steps of resolution theorem-proving, which relies on the method of *reductio ad absurdum*, i.e. showing that the negation of the desired conclusion leads to a contradiction. Unfortunately, despite the use of various heuristics to speed up the search, the method still remains intractable: the search space tends to

Table 2.1
A simple example of "resolution" theorem-proving

The deduction to be evaluated:
1. Mary is a psychologist.
2. All psychologists have read some books.
3. ∴ Mary has read some books.

Step 1: Translate the deduction into a *reductio ad absurdum*, i.e. negate the conclusion with the aim of showing that the resultant set of propositions is inconsistent:
1. (Psychologist Mary)
2. (For any x)(for some y)
   ((Psychologist x) → ((Book y) & (Read x y)))
3. (Not (For some z) (Book z & (Read Mary z)))

Step 2: Translate all the connectives into disjunctions, and eliminate the quantifiers. "Any" can be deleted: its work is done by the presence of variables. "Some" is replaced by a function (the so-called Skolem function), e.g. "all psychologists have read some books" requires a function, f, which, given a psychologist as its argument, returns a value consisting of some books:
1. (Psychologist Mary)
2. (Not (Psychologist x)) or (Read x (f x))
3. (Not (Read Mary (f Mary))

Step 3: Apply the resolution rule to any premises containing inconsistent clauses: it is not necessary for both assertions to be disjunctions. Assertion 3 thus cancels out the second disjunct in assertion 2 to leave:
1. (Psychologist Mary)
2. (not (Psychologist Mary))
These two assertions cancel out by a further application of the resolution rule. Whenever a set of assertions is reduced to the empty set in this way, they are inconsistent. The desired conclusion follows at once because its negation has led to a *reductio ad absurdum*.

grow exponentially with the number of clauses in the premises (Moore, 1982). The resolution method, however, has become part of "logic programming"—the formulation of high level programming languages in which programs consist of assertions in a formalism closely resembling the predicate calculus (Kowalski, 1979). Thus, the language PROLOG is based on resolution (see e.g. Clocksin and Mellish, 1981).

No psychologist would suppose that human reasoners are equipped with the resolution rule (see also our studies of "double disjunctions" in the next chapter). But, a psychologically more plausible form of

deduction has been implemented in computer programs. It relies on the method of "natural deduction", which we described in Chapter 1, and which provides separate rules of inference for each connective. The programs maintain a clear distinction between what has been proved and what their goals are, and so they are able to construct chains of inference working forwards from the premises and working backwards from the conclusion to be proved (see e.g. Reiter, 1973; Bledsoe, 1977; Pollock, 1989). The use of forward and backward chains was pioneered in modern times by Polya (1957) and by Newell, Shaw, and Simon (1963); as we will see, it is part of the programming language, PLANNER.

Formal Rules in Psychological Theories

Natural deduction has been advocated as the most plausible account of mental logic by many psychologists (e.g. Braine, 1978; Osherson, 1975; Johnson-Laird, 1975; Macnamara, 1986), and at least one simulation program uses it for both forward- and backward-chaining (Rips, 1983). All of these theories posit an initial process of recovering the logical form of the premises. Indeed, what they have in common outweighs their differences, but we will outline three of them to enable readers to make up their own minds.

Johnson-Laird (1975) proposed a theory of propositional reasoning partly based on natural deduction. Its rules are summarized in Table 2.2 along with those of the two other theories. The rule introducing disjunctive conclusions:

A
∴ A or B (or both)

leads to deductions that, as we have remarked, throw semantic information away and thus seem unacceptable to many people. Yet, without this rule, it would be difficult to make the inference:

If it is frosty or it is foggy, then the game won't be played.
It is frosty.
Therefore, the game won't be played.

Johnson-Laird therefore proposed that the rule (and others like it) is an auxiliary one that can be used only to prepare the way for a primary rule, such as modus ponens. Where the procedures for exploiting rules fail, then the next step, according to his theory, is to make a hypothetical assumption and to follow up its consequences.

Braine and his colleagues have described a series of formal theories based on natural deduction (see e.g. Braine, 1978; Braine and Rumain, 1983). At the heart of their approach are the formal rules presented in Table 2.2. They differ in format from Johnson-Laird's in two ways. First, "and" and "or" can connect any number of propositions, and so, for example, the first rule in Table 2.2 has the following form in their theory:

$P_1, P_2, \ldots P_n$
Therefore, P_1 and P_2 and $\ldots P_n$.

Second, Braine avoids the need for some auxiliary rules, such as the disjunctive rule above, by building their effects directly into the main rules. He includes, for example, the rule:

If A or B then C
A
Therefore C

again allowing for any number of propositions in the disjunctive antecedent. This idea is also adopted by Sperber and Wilson (1986).

Braine, Reiser, and Rumain (1984) tested the theory by asking subjects to evaluate given deductions. The problems concerned the presence or absence of letters on an imaginary blackboard, e.g.:

If there is either a C or an H, then there is a P.
There is a C.
Therefore, there is a P.

The subjects' task was to judge the truth of the conclusion given the premises. The study examined two potential indices of difficulty—the number of steps in a deduction according to the theory, and the "difficulty weights" of these steps as estimated from the data. Both measures predicted certain results: the rated difficulty of a problem, the latency of response (adjusted for the time it took to read the problem), and the percentage of errors. Likewise, the number of words in a problem correlated with its rated difficulty and the latency of response.

Rips (1983) has proposed a theory of propositional reasoning, which he has simulated in a program called ANDS (A Natural Deduction System). The rules used by the program—in the form of procedures—are summarized in Table 2.2. The program evaluates given conclusions and it builds both forward-chains and backward-chains of deduction, and therefore maintains a set of goals separate from the assertions that it

Table 2.2
The principal formal rules of inference proposed by three psychological theories of deduction

	Johnson-Laird	*Braine*	*Rips*
Conjunctions			
A, B ∴ A & B	+	+	+
A & B ∴ A	+	+	+
Disjunctions			
A or B, not-A ∴ B	+	+	+
A ∴ A or B	+		+
Conditionals			
If A then B, A ∴ B	+	+	+
If A or B then C, A ∴ C		+	+
A ⊢ B ∴ If A then B	+	+	+
Negated conjunctions			
not (A & B), A ∴ ¬B	+	+	
not (A & B) ∴ not-A or not-B			+
A & not-B ∴ not (A & B)	+		
Double negations			
not not-A ∴ A	+	+	
De Morgan's laws			
A & (B or C) ∴ (A & B) or (A & C)		+	
Reductio ad absurdum			
A ⊢ B & not-B ∴ not-A	+	+	+
Dilemmas			
A or B, A ⊢ C, B ⊢ C ∴ C		+	+
A or B, A ⊢ C, B ⊢ D ∴ C or D		+	
Introduction of tautologies			
∴ A or not-A		+	+

Notes
"+" indicates that a rule is postulated by the relevant theory.
"A ⊢ B" means that a deduction from A to B is possible. Braine's rules interconnect any number of propositions, as we explain in the text. He postulates four separate rules that together enable a *reductio ad absurdum* to be made. Johnson-Laird relies on procedures that follow up the separate consequences of constituents in order to carry out dilemmas.

has derived. Certain rules are treated as auxiliaries that can be used only when they are triggered by a goal, e.g.:

A, B
Therefore, A and B

which otherwise could be used *ad infinitum* at any point in the proof. If the program can find no rule to apply during a proof, then it declares that the argument is invalid. Rips assumes that rules of inference are available to human reasoners on a probabilistic basis. His main method of testing the theory has been to fit it to data obtained from subjects who assessed the validity of arguments. The resulting estimates of the availability of rules yielded a reasonable fit for the data as a whole. One surprise, however, was that the rule:

If A or B then C
A
Therefore, C

had a higher availability than the simple rule of modus ponens. It is worth noting that half of the valid deductions in his experiment called for semantic information to be thrown away. Only one out of these 16 problems was evaluated better than chance. Conversely, 14 of the other 16 problems, which maintained semantic information, were evaluated better than chance.

A major difficulty for performance theories based on formal logic is that people are affected by the content of a deductive problem. We will discuss a celebrated demonstration of this phenomenon—Wason's selection task—in Chapter 4. Yet, formal rules ought to apply regardless of content. That is what they are: rules that apply to the logical form of assertions, once it has been abstracted from their content. The proponents of formal rules argue that content exerts its influence only during the interpretation of premises. It leads reasoners to import additional information, or to assign a different logical form to a premise. A radical alternative, however, is that reasoners make use of rules of inference that have a specific content.

CONTENT-SPECIFIC RULES: A SECOND THEORY AT THE ALGORITHMIC LEVEL

Content-specific rules of inference were pioneered by workers in artificial intelligence. They were originally implemented in the programming language PLANNER (Hewitt, 1971). It and its many descendants rely on the resemblance between proofs and plans. A proof

is a series of assertions, each following from what has gone before, that leads to a conclusion. A plan is a series of hypothetical actions, each made possible by what has gone before, and leading to a goal. Hence, a plan can be derived in much the same way as a proof. A program written in a PLANNER-like language has a data-base consisting of a set of simple assertions, such as:

Mary is a psychologist.
Paul is a linguist.
Mark is a programmer.

which can be represented in the following notation:

(Psychologist Mary)
(Linguist Paul)
(Programmer Mark)

The assertion, "Mary is a psychologist", is obviously true with respect to this data base. General assertions, such as:

All psychologists are experimenters.

are expressed, not as assertions, but as rules of inference. One way to formulate such a rule is by a procedure:

(Consequent (x) (Experimenter x)
(Goal (Psychologist x)))

which enables the program to infer the consequent that x is an experimenter if it can satisfy the goal that x is a psychologist. If the program has to evaluate the truth of:

Mary is an experimenter

it first searches its data base for a specific assertion to that effect. It fails to find such an assertion in the data base above, and so it looks for a rule with a consequent that matches with the sentence to be evaluated. The rule above matches and sets up the following goal:

(Goal (Psychologist Mary))

This goal *is* satisfied by an assertion in the data base, and so the sentence, "Mary is an experimenter" is satisfied too. The program

constructs backward-chains of inference using such rules, which can even be supplemented with specific heuristic advice about how to derive certain conclusions.

Another way in which to formulate a content-specific rule is as follows:

(Antecedent (x) (Psychologist x)
(Assert (x)(Experimenter x)))

Whenever its antecedent is satisfied by an input assertion, such as:

Mary is a psychologist.

the procedure springs to life and asserts that x is an experimenter:

Mary is an experimenter.

This response has the effect of adding the further assertion to the data base. The program can construct forward-chains of inference using such rules.

Content-specific rules are the basis of most expert systems, which are computer programs that give advice on such matters as medical diagnosis, the structure of molecules, and where to drill for minerals. They contain a large number of conditional rules that have been culled from human experts. From a logical standpoint, these rules are postulates that capture a body of knowledge. The expert systems, however, use them as rules of inference (see e.g. Michie, 1979; Duda, Gaschnig, and Hart, 1979; Feigenbaum and McCorduck, 1984). The rules are highly specific. For example, DENDRAL, which analyzes mass spectrograms (Lindsay, Buchanan, Feigenbaum, and Lederberg, 1980), includes this conditional rule:

If there is a high peak at 71 atomic mass units
and there is a high peak at 43 atomic mass units
and there is a high peak at 86 atomic mass units
and there is any peak at 58 atomic mass units
then there must be an N-PROPYL-KETONE3 substructure.

(see Winston, 1984, p.196). Most current systems have an inferential "engine" which, by interrogating a user about a particular problem, navigates its way through the rules to yield a conclusion. The conditional rules may be definitive or else have probabilities associated with them, and the system may even use Bayes theorem from the probability

calculus. It may build forward chains (Feigenbaum, Buchanan, and Lederberg, 1979), backward chains (Shortliffe, 1976), or a mixture of both (Waterman and Hayes-Roth, 1978).

Psychologists have also proposed that the mind uses content-specific conditional rules to represent general knowledge (e.g. Anderson, 1983). They are a plausible way of drawing inferences that depend on background assumptions. The proposal is even part of a seminal theory of cognitive architecture in which the rules (or "productions"as they are known) are triggered by the current contents of working memory (see Newell and Simon, 1972, and Newell, 1990). When a production is triggered it may, in turn, add new information to working memory, and in this way a chain of inferences can ensue.

A variant on content-specific rules has been proposed by Cheng and Holyoak (1985), who argue that people are guided by "pragmatic reasoning schemas." These are general principles that apply to a particular domain. For example, there is supposedly a permission schema that includes rules of the following sort:

If action A is to be taken then precondition B must be satisfied.

The schema is intended to govern actions that occur within a framework of moral conventions, and Cheng and Holyoak argue that it and other similar schemas account for certain aspects of deductive performance (see Chapter 4).

Content plays its most specific role in the hypothesis that reasoning is based on memories of particular experiences (Stanfill and Waltz, 1986). Indeed, according to Riesbeck and Schank's (1989) theory of "case-based" reasoning, human thinking has nothing to do with logic. What happens is that a problem reminds you of a previous case, and you decide what to do on the basis of this case. These theorists allow, however, that when an activity has been repeated often enough, it begins to function like a content-specific rule. The only difficulty with this theory is that it fails to explain how people are able to make valid deductions that do not depend on their specific experiences.

General knowledge certainly enters into everyday deductions, but whether it is represented by schemas or productions or specific cases is an open question. It might, after all, be represented by *assertions* in a mental language. It might even have a distributed representation that has no explicit symbolic structure (Rumelhart, 1989). Structured representations, however, do appear to be needed in order to account for reasoning about reasoning (see Chapter 9, and Johnson-Laird, 1988b, Chapter 19).

MENTAL MODELS: A THIRD THEORY AT THE ALGORITHMIC LEVEL

Neither formal rules nor content-specific rules appear to give complete explanations of the mechanism underlying deduction. On the one hand, the content of premises can exert a profound effect on the conclusions that people draw, and so a uniform procedure for extracting logical form and applying formal rules to it may not account for all aspects of performance. On the other hand, ordinary individuals are able to make valid deductions that depend solely on connectives and quantifiers, and so rules with a specific content would have to rely on some (yet to be formulated) account of purely logical competence. One way out of this dilemma is provided by a third sort of algorithmic theory, which depends on semantic procedures.

Consider this inference:

The black ball is directly behind the cue ball. The green ball is on the right of the cue ball, and there is a red ball between them.
Therefore, if I move so that the red ball is between me and the black ball, the cue ball is to the left of my line of sight.

It is possible to frame rules that capture this inference (from Johnson-Laird, 1975), but it seems likely that people will make it by imagining the layout of the balls. This idea lies at the heart of the theory of mental models. According to this theory, the process of deduction depends on three stages of thought, which are summarized in Figure 2.1. In the first stage, comprehension, reasoners use their knowledge of the language and their general knowledge to understand the premises: they construct an internal model of the state of affairs that the premises describe. A deduction may also depend on perception, and thus on a perceptually-based model of the world (see Marr, 1982). In the second stage, reasoners try to formulate a parsimonious description of the models they have constructed. This description should assert something that is not explicitly stated in the premises. Where there is no such conclusion, then they respond that nothing follows from the premises. In the third stage, reasoners search for alternative models of the premises in which their putative conclusion is false. If there is no such model, then the conclusion is valid. If there is such a model, then prudent reasoners will return to the second stage to try to discover whether there is any conclusion true in all the models that they have so far constructed. If so, then it is necessary to search for counterexamples to it, and so on, until the set of possible models has been exhausted. Because the number

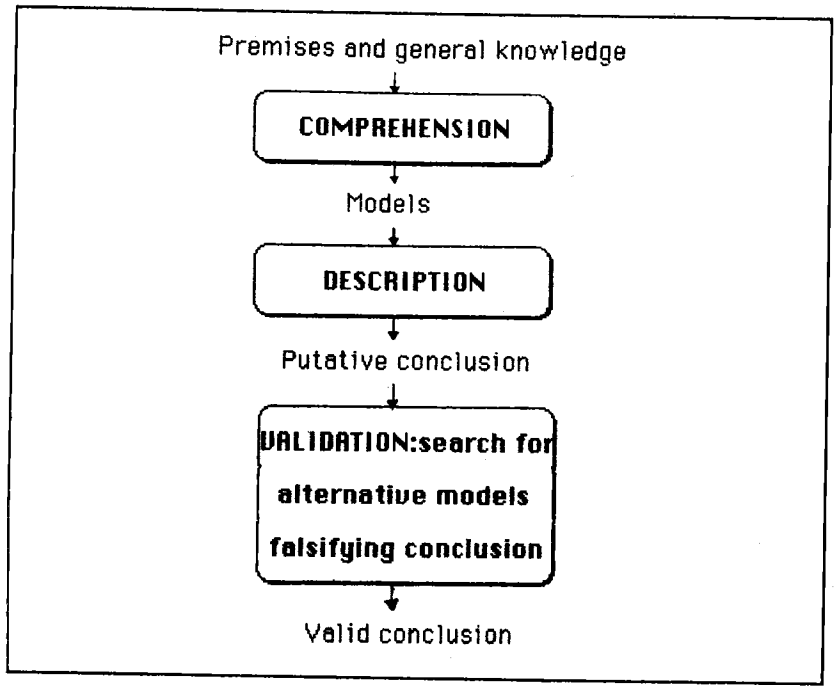


Figure 2.1. The three stages of deduction according to the model theory.

of possible mental models is finite for deductions that depend on quantifiers and connectives, the search can in principle be exhaustive. If it is uncertain whether there is an alternative model of the premises, then the conclusion can be drawn in a tentative or probabilistic way. Only in the third stage is any essential deductive work carried out: the first two stages are merely normal processes of comprehension and description.

The theory is compatible with the way in which logicians formulate a semantics for a calculus (see Chapter 1). But, logical accounts depend on assigning an infinite number of models to each proposition, and an infinite set is far too big to fit inside anyone's head (Partee, 1979). The psychological theory therefore assumes that people construct a minimum of models: they try to work with just a single representative sample from the set of possible models, until they are forced to consider alternatives.

Models form the basis of various theories of reasoning. An early program for proving geometric theorems used diagrams of figures in order to rule out subgoals that were false (Gelernter, 1963). Although

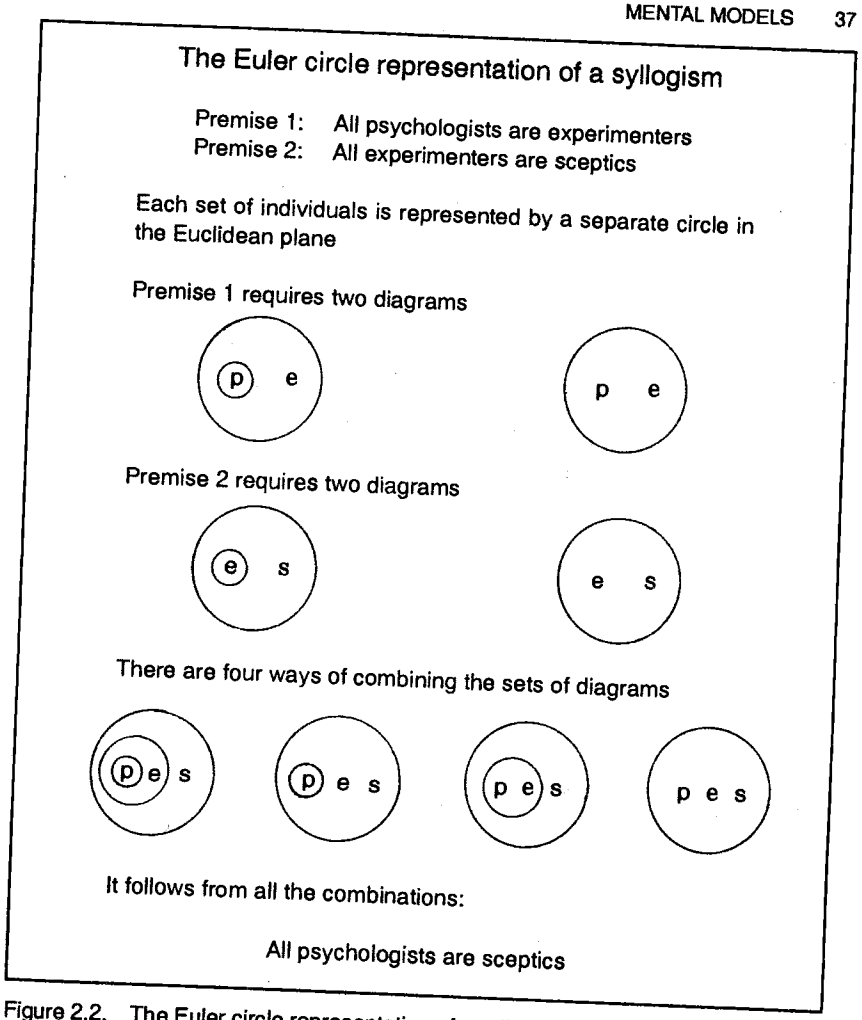


Figure 2.2. The Euler circle representation of a syllogism.

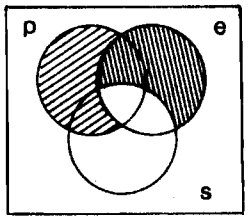
this idea could be used in other domains (see Bundy, 1983), there have been few such applications in artificial intelligence. Charniak and McDermott (1985, p.363) speculate that the reason might be because few domains have counterexamples in the form of diagrams. Yet, as we will see, analogous structures are available for all sorts of deduction.

Deductions from singly-quantified premises, such as "All psychologists are experimenters", can be modelled using Euler circles (see Figure 2.2). Psychological theories have postulated such representations (Erickson, 1974) or equivalent strings of symbols

The Venn diagram representation of a syllogism

Premise 1: All psychologists are experimenters
Premise 2: All experimenters are sceptics

Each of the three sets is initially represented by one of three overlapping circles within a rectangle that represents the universe of discourse.



Premise 1 rules out the possibility of psychologists who are not experimenters, and so the corresponding portion of the circle representing psychologists is shaded out.
Premise 2 likewise rules out the possibility of experimenters who are not sceptics, and so the corresponding portion of the circle representing experimenters is shaded out. The resulting diagram establishes the conclusion:

All psychologists are sceptics.

Figure 2.3. The Venn diagram representation of a syllogism.

(Guyote and Sternberg, 1981). These deductions can also be modelled using Venn diagrams (see Figure 2.3) or equivalent strings of symbols, and they too have been proposed as mental representations (Newell, 1981). A uniform and more powerful principle, however, is that *mental models have the same structure as human conceptions of the situations they represent* (Johnson-Laird, 1983). Hence, a finite set of individuals is represented, not by a circle inscribed in Euclidean space, but by a finite set of mental tokens. A similar notion of a "vivid" representation has been proposed by Levesque (1986) from the standpoint of developing efficient computer programs for reasoning. But, there are distinctions between the two sorts of representation, e.g. vivid representations cannot represent directly either negatives or disjunctions (see also

Etherington, Borgida, Brachman, and Kautz, 1989). The tokens of mental models may occur in a visual image, or they may not be directly accessible to consciousness. What matters is, not the phenomenal experience, but the structure of the models. This structure, which we will examine in detail in the following chapters, often transcends the perceptible. It can represent negation and disjunction.

The general theory of mental models has been successful in accounting for patterns of performance in various sorts of reasoning (Johnson-Laird, 1983). Errors occur, according to the theory, because people fail to consider all possible models of the premises. They therefore fail to find counterexamples to the conclusions that they derive from their initial models, perhaps because of the limited processing capacity of working memory (Baddeley, 1986).

The model theory has attracted considerable criticism from adherents of formal rules. It has been accused of being unclear, unworkable, and unnecessary. We will defer our main reply to critics until the final chapter, but we will make a preliminary response here to the three main charges that the theory is empirically inadequate:

1. Mental models do not explain propositional reasoning: "No clear mental model theory of propositional reasoning has yet been proposed" (Braine et al., 1984; see also Evans, 1984, 1987; and Rips, 1986). The next chapter renders this criticism obsolete.

2. Mental models cannot account for performance in Wason's selection task. The theory implies that people search for counterexamples, yet they conspicuously fail to do so in the selection task (Evans, 1987). The criticism is based on a false assumption. The theory does not postulate that the search for counterexamples is invariably complete—far from it, as such an impeccable performance would be incompatible with observed errors. In Chapter 4, we will show how the theory explains performance in the selection task.

3. Contrary to the previous criticism, Rips (1986) asserts: "Deduction-as-simulation explains content effects, but unfortunately it does so at the cost of being unable to explain the generality of inference". He argues that a modus ponens deduction is not affected by the complexity of its content, and is readily carried out in domains for which the reasoner has had no previous exposure and thus no model to employ. However, the notion that reasoners cannot construct models for unfamiliar domains is false: all they need is a knowledge of the meaning of the connectives and other logical terms that occur in the premises. Conversely, modus ponens can be affected by its content as we will show in Chapter 4.

CONCLUSION

We have completed our survey of where things stood at the start of our research. There were—and remain—three algorithmic theories of deduction. Despite many empirical findings, it had proved impossible to make a definitive choice among the theories. We now turn to the studies that will enable us to reach an informed decision about their adequacy as accounts of human deductive performance.