

**Copyright** © 1991 by Lawrence **Erlbaum** Associates Ltd.  
All rights reserved. No part of this book may be reproduced in any form, by **photostat**, microform, retrieval system, or any other means without the prior written permission of the publisher.

Reprinted 1992

Lawrence Erlbaum Associates Ltd., Publishers  
27 **Palmeira** Mansions  
Church Road  
Hove  
East Sussex, BN3 2FA  
U.K.

### **British Library Cataloguing in Publication Data**

Johnson-Laird, P.N. (Philip Nicholas)

I. Title II. Byrne, Ruth **M.J.**

162

Librarians please shelve under subject classification:  
Psychology. Human Deduction

ISBN 0-86377-148-3 (Hbk)

ISBN 0-86377-149-1 (Pbk)

ISSN 0959-4779 (Essays in Cognitive Psychology)

# Contents

**Prologue** ix

**Acknowledgements** xi

**Chapter 1: The Logic of Deduction** 1

Introduction and plan of the book 1

The concept of logical form 4

The propositional calculus 5

The predicate calculus 11

**Chapter 2: The Cognitive Science of Deduction** 17

Deduction: A theory at the computational level 18

Formal rules: A theory at the algorithmic level 23

Content-specific rules: A second theory at the algorithmic level 31

Mental models: A third theory at the algorithmic level 35

Conclusion 40

**Chapter 3: Reasoning with Propositions** 41

Models for connectives 43

Three phenomena predicted by the model theory 52

Conclusions 61

**Chapter 4: Conditionals** 63

The model theory of the meaning of conditionals 65

Deduction with conditionals 73

Conclusions 85

# 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)



# The Logic of Deduction

## INTRODUCTION AND PLAN OF THE BOOK

Sherlock Holmes popularized a profound misconception about deduction. At their first meeting, the great detective surprised Dr. Watson by remarking: "you have been in **Afghanistan**". Later he explained his methods:

From long habit the train of thoughts ran so swiftly through my mind that I arrived at the conclusion without being **conscious** of intermediate steps. There were such **steps**, however. The train of reasoning ran, "Here is a gentleman of a medical type, but with the air of a military man. Clearly, an army doctor, then. He has just come from the tropics, for his face is dark, and that is not the natural tint of his skin for his wrists are fair. He has undergone hardship and sickness, as his haggard face says clearly. His left arm has been injured. He holds it in a stiff and unnatural manner. Where in the tropics could an English army doctor have seen much hardship and got his arm wounded? Clearly in Afghanistan." **The whole train of thought did not occupy a second. I then remarked that you came from Afghanistan, and you were astonished.**

(Conan Doyle, 1892, p.24)

Holmes is undoubtedly reasoning, but is he making deductions? Granted that his perceptions and background knowledge are accurate, does it follow that his conclusion must be true? Of course, not. He could have blundered, and Watson might have replied:

I am an army doctor, but I have *not* been in Afghanistan. I have been in a Swiss Sanatorium recovering from TB. The sun is responsible for my tan, and my arm was injured in a climbing accident.

Holmes reached a plausible conclusion but he did not make a valid deduction. By definition, a *valid* deduction yields a conclusion that must be true given that its premises are true. The great detective did not always succeed in his cases, but remarkably he seems never to have drawn a conclusion that turned out to be false.

Our topic is deduction, but the case of Sherlock Holmes forces us to consider other sorts of thinking if only to put deduction in its proper place. Plausible inference is **different** from deduction; daydreaming is different from problem solving; mental arithmetic is different from making a decision. But intuition soon ceases to be a reliable guide to the varieties of thought. A more systematic taxonomy can be based on a computational analysis, and it yields several main sorts of thought (see Johnson-Laird, 1988a). A process of thought can be directed by a goal or it may flow undirected in a "stream of **consciousness**" in which each idea suggests another. Hobbes (1651) drew the same distinction:

This **Trayne** of Thoughts, or **Mentall** Discourse, is of two sorts. The first is *Unguided*, *without Designe*, and inconstant .... In which case the thoughts are said to wander, and seem impertinent one to another, as in a **Dream**.... The second is more constant; as being *regulated* by some desire, and **designe**. (p. 95)

Among **goal-directed** thinking, reasoning begins with a definite starting **point**—a set of observations or premises, and so it can be distinguished from creative processes, which can occur with no clear starting point beyond, say, a blank canvas or an empty sheet of paper. There are three main varieties of reasoning: calculation, deduction, and induction. Calculation is the routine application of a procedure known by heart, as in mental arithmetic. Deduction is a less systematic process in which the goal is to draw a valid consequence from premises. Induction sacrifices validity for plausibility. Like Sherlock Holmes, one often does not have sufficient information to be able to draw a valid inference.

Association, creation, induction, deduction, and calculation, underlie all forms of thought, and so a complete theory of thinking has to explain each of them. In this monograph, we have set ourselves a more modest goal: to explain the nature of deduction and to characterize its underlying mental processes.

Why deduction? One reason is because of its intrinsic importance: it plays a crucial role in many tasks. You need to make deductions in order

to formulate plans and to evaluate actions; to determine the consequences of assumptions and hypotheses; to interpret and to formulate instructions, rules, and general principles; to pursue arguments and negotiations; to weigh evidence and to assess data; to decide between competing theories; and to solve problems. A world without deduction would be a world without science, technology, laws, social conventions, and culture. And if you want to dispute this claim, we shall need to assess the validity of your arguments. Another reason for studying deduction is that it is ripe for solution. Unlike the other processes of thought, deduction has been studied sufficiently to be within our grasp. Psychologists have accumulated 80 years\* worth of experiments on deductive reasoning (for reviews, see e.g. Wason and Johnson-Laird, 1972; Evans, 1982); they have proposed explicit information-processing theories of its mechanism (see e.g., Erickson, 1974; Braine, 1978; Johnson-Laird, 1983); and workers in artificial intelligence have developed many computer programs that carry out deduction (see e.g. Reiter, 1973; Doyle, 1979; Robinson, 1979; **McDermott**, 1987).

The present book focuses on our own research, and its plan is simple. In this chapter, we provide a brief but necessary background in logic. We then plunge into the murkier problems of psychology, beginning with the theories of reasoning that were current when we started our research in October 1986. The **major** part of the book is devoted to our own studies. We start with deductions that depend on propositions that have been formed using such connectives as “**and**”, “**or**”, and “**not**”. We then spend an entire chapter on the **notorious** problems of “**if**”—**problems** that have led some philosophers to abandon hope of a semantic analysis of conditional assertions. Next, we consider deductions that depend on relations between entities, such as “Lisa is taller than **Jenny**”, “Steven is in the same place as Paul”, or “Cathy is related to Linda”. Relations can hold between sets of individuals and, hence deductions about them often hinge on such words as “**any**”, and “**some**”, and so we consider these quantifiers. We deal with arguments in which each premise contains only a single quantifier (i.e. “**sylogisms**”), before we turn to premises containing more than one quantifier, such as, “none of the children is in the same place as any of the adults”. The three domains of **propositional**, relational, and **quantificational** reasoning exhaust the main sorts of deduction. But, an important higher-order sort of thinking is *meta-deduction*, in which deductions concern other deductions or statements explicitly assigning truth or falsity to **propositions**. **Meta-deduction** is an important ability because without it human beings are unlikely to have invented formal logic as a discipline. We therefore devote a chapter to the topic. In

developing our theory, we have written computer programs that model it. In the penultimate chapter, we describe how to build programs that make deductions from models. We address a problem that workers in artificial intelligence have often confronted: “**non-monotonic**” reasoning, which is the withdrawal of a conclusion in the light of subsequent information. We also address a problem that they have often evaded: how to draw conclusions that are maximally parsimonious. This work solves a **major** problem in the design of electronic circuits: how to find the simplest possible circuit for carrying out a particular Boolean operation. Finally, we recapitulate our theory of human reasoning, explore its consequences for cognitive science, and attempt to answer our critics.

We begin with logic because it is the true science of deduction (*pace* Holmes), which arose from a need to determine whether or not inferences are valid. We give a brief introduction to the subject in order to establish the distinction between formal and semantic methods. Readers should proceed at once to the next chapter if they understand why there are direct semantic methods for testing validity in the **propositional** calculus, but *not* in the predicate calculus.

### THE CONCEPT OF LOGICAL FORM

Aristotle, who by his own account was the first logician, noted that certain inferences are valid in virtue of their *form* (see e.g. Kneale and Kneale, 1962). Thus, the valid argument:

All cows are mammals.  
 All mammals are warm-blooded.  
 Therefore, all cows are warm-blooded.

has the form:

—  
**All A are B.**  
**All B are C.**  
 Therefore, **All A are C.**

No matter what terms we substitute for A, B, and C, the result is a valid deduction, e.g.:

All politicians are authoritarians.  
 All authoritarians are virtuous.  
 Therefore, all politicians are virtuous.

You may object that this last conclusion is patently false. And so it is, but remember that a valid conclusion is one that must be true *if* its premises are true. In this case, the premises are false; had they been true, the conclusion would have been true too. Aristotle based his logic on arguments of this sort, which are known as *sylogisms*, and which were long thought to be the centre of the subject; we now know that they are one of its minor suburbs. Nevertheless, the notion of the *form* of a deduction has been central to the development of logic.

Form is a matter of syntax: it depends on the position of certain words, such as "all" and "some", and of other terms within the premises and conclusion. Hence, formal logic is in essence a syntactic device for testing whether the form of an argument is a valid one. Indeed, Leibniz (1666) dreamt of a universal system that would enable all disputes to be resolved by such dispassionate calculations. A step towards the realization of this dream was the invention of the propositional calculus in the nineteenth century.

## THE PROPOSITIONAL CALCULUS

The propositional calculus deals with arguments that depend on sentences containing such connectives as "not", "if", "and", and "or". It can be set up in many different, though equivalent, ways. One way, the method of "natural deduction", uses formal rules of inference for each of the different connectives (see e.g. Gentzen, 1935; Prawitz, 1965). A typical example of a formal rule is the following one, which is known as the rule of *modus ponens*:

If p then q  
 p  
 Therefore, q.

where p and q are variables that can denote any propositions, no matter how complex. The rule can be used to make the deduction:

If Arthur is in Edinburgh, then Carol is in Glasgow.  
 Arthur is in Edinburgh.  
 Therefore, Carol is in Glasgow.

The premises have a logical form that matches the rule, where p equals "Arthur is in Edinburgh" and q equals "Carol is in Glasgow", and so it yields the conclusion. In general, a conclusion can be derived from premises provided that it has a proof in which each step can be made by

applying a formal rule of inference to one or more of the earlier assertions in the derivation. Logicians refer to this formal method as “**proof-theoretic**”.

What about the meanings of the connectives? In fact, they can be given an explicit account that motivates the choice of formal rules. Each proposition is assumed to be either true or false. Strictly speaking, it is a mistake to assign truth or falsity to *sentences* in a natural language, because a sentence can be used to assert many different propositions, e.g. “I am here” asserts **different** propositions depending on who asserts it and where they are. Hence, it is propositions, not sentences, that are true or false (see Strawson, 1950). The meaning of a connective, such as “**and**”, can then be defined solely as a function of the truth values of the propositions that it interconnects. Thus, the conjunction of any two propositions:

**p and q**

is itself true provided that both of its two constituent propositions are true, and it is false only if one or both of them are false. This definition can be stated in a truth table:

p	q	p and q
True	True	True
True	False	False
False	True	False
False	False	False

where each row represents a separate possible combination of truth values. Because by assumption any proposition is either true or false, there are only four rows in a truth table based on two propositions. The truth value of the conjunction depends solely on the truth values of p and q, and, as the table shows, it is true only if both of them **are** true.

The meaning of “or” is defined analogously. An *inclusive* disjunction of any two propositions:

**p or q, or both**

is true provided that at least one of its two constituent propositions is true, and it is false only if they are both false. This definition can also be stated in a truth table:

p	q	p or q (or both)
True	True	True
True	False	True
False	True	True
False	False	False

An *exclusive* disjunction:

p or q, but not both

in contrast, is true only when one of the two propositions is true.

In ordinary discourse, a connective such as “and” may transcend a truth-table definition. The assertion:

He fell **off** his bicycle and broke his leg.

is usually taken to mean that the first event occurred before the second. Hence, the following conjunction has a different meaning:

He broke his leg and fell off his bicycle.

The two assertions are synonymous according to the truth table definition. This divergence has led some philosophers to argue that the meaning of “and” in certain of its uses does not correspond to any truth table: it can express a temporal or causal relation. Others, notably Grice (1975), have defended the truth table, and argued that it is merely overlaid by pragmatic factors that depend on general knowledge and the conventions of discourse.

The most puzzling case is that of conditional assertions, such as:

If Arthur is in Edinburgh, then Carol is in Glasgow.

This proposition is true when its antecedent (“Arthur is in Edinburgh”) and consequent (“Carol is in Glasgow”) are both true, and false when its antecedent is true and its consequent false. But, suppose that its antecedent is false, i.e. Arthur is *not* in Edinburgh, is the conditional then true or false? It can hardly be false, and so, since the propositional calculus allows only truth or falsity, it must be true. This treatment yields the following truth table, and the corresponding connective in the calculus is known as *material implication*:

p	q	if p then q
T	T	T
T	F	F
F	T	T
F	F	T

where we have abbreviated "True" as "T" and "False" as "F".

In everyday language, conditional assertions often suggest a relation between antecedent and consequent (see Fillenbaum, 1977). The relation may be a cause or a reason, e.g. "If you tidy your room, then I will take you to the cinema", conveys an implicit negative relation, "if you don't tidy your room, then I won't take you to the cinema". The positive and negative relations taken together are equivalent to the **bi-conditional**:

If *and only if* you tidy your room, then I will take you to the cinema.

This connective has the following truth table, and it is known as *material equivalence*:

p	q	if and only if p then q
T	T	T
T	F	F
F	T	F
F	F	T

The meaning of "not" concerns only a single proposition, and it reverses its truth value:

P	not p
T	F
F	T

**Logicians** have considered the equivalences between different combinations of connectives, and they have shown that any truth table can be expressed in terms of negation and inclusive disjunction. For example, the truth table for material implication corresponds to the inclusive disjunction, not-p or q. This construal rather than the conditional, if p then q, perhaps yields a better correspondence between language and logic.

We have shown how to derive conclusions from premises using formal rules, which are sensitive only to the syntactic form of expressions. The reader may be wondering whether there is a way to make deductions that depends, not on the form of expressions, but on their meaning. There is a way that we will now illustrate using the following premises:

Arthur is in Edinburgh or Betty is in Dundee, or both.  
 Betty is not in Dundee.  
 If Arthur is in Edinburgh, then Carol is in Glasgow.

We abbreviate the individual atomic propositions as follows: a for "Arthur is in Edinburgh", b for "Betty is in Dundee", and c for "Carol is in Glasgow". The set of possibilities for these three propositions is:

a	b	c
T	T	T
T	T	F
T	F	T
T	F	F
F	T	T
F	T	F
F	F	T
F	F	F

What we want to discover is which of these possibilities, if any, must be true given the truth of the premises. The first premise, a or b (or both), rules out two possibilities, i.e. those in which a and b are both false:

a	b	c	a or b, or both
T	T	T	
T	T	F	
T	F	T	
T	F	F	
F	T	T	
F	T	F	
F	F	T	F
F	F	F	F

The second premise, not-b, rules out four further possibilities, i.e. those in which b is true:

a	b	c	not-b
T	T	T	F
T	T	F	F
T	F	T	
T	F	F	
F	T	T	F
F	T	F	F

The third premise, if a then c, eliminates the case in which a is true and c is false:

a	b	c	if a then c
T	F	T	
T	F	F	F

As Sherlock Holmes remarked, when you have eliminated the impossible then whatever remains, however improbable, must be the case. There remains only a single possibility, in which c is true, and so the corresponding proposition:

Carol is in Glasgow.

is a valid conclusion. In general, a conclusion can be derived from premises provided that the premises eliminate all but the contingencies in which it is true. Logicians refer to this semantic method as "model-theoretic".

We have demonstrated two distinct ways of making deductions in the **propositional** calculus. The "proof-theoretic" way is syntactic and depends on formal rules of inference. The "model-theoretic" way is **semantic** and depends on eliminating those states of affairs that are false given the truth of the premises.

In the case of the **propositional** calculus, it can be proved that if a conclusion is valid using the semantic method then it can be derived using the syntactic method, and the calculus is therefore said to be "complete". Conversely, if a conclusion can be derived using the syntactic method, then it is valid using the semantic method, and the calculus is therefore said to be "sound". You might therefore imagine that the two methods are merely trivial variants of one another. You would be wrong. The higher-order predicate calculus (to be described presently) is not complete: no consistent formalization of it can capture all valid conclusions. What is valid cannot always be established by syntactic derivations from the logical form of premises. This mismatch drives a

wedge between syntax and semantics. They are not trivial variants of one another.

## THE PREDICATE CALCULUS

Many deductions in daily life hinge on matters internal to propositions rather than on the external connections between them. For instance, the *relational* deduction:

Anna is in the same place as Ben.  
Ben is in the same place as Con.  
Therefore, Anna is in the same place as Con.

is intuitively valid, yet its validity cannot be accounted for in the *propositional* calculus, for which the argument has the form:

**p**  
**q**  
Therefore, **r**.

What we need is the predicate calculus, which includes the *propositional* calculus as a proper part, but which goes beyond it by introducing machinery for dealing with the internal structure of propositions. Each of the premises in the example contains the relation, "is in the same place as", which for convenience we will abbreviate as follows:

Anna in-same-place Ben.  
Ben in-same-place Con.

Before the deduction can proceed using formal rules, we have to introduce a further premise that expresses the fact that the relation is transitive, which we can express in "Loglish", a language that closely resembles the predicate calculus, as:

For any  $x$ , any  $y$ , any  $z$ , if  $x$  in-same-place  $y$ , and  $y$  in-same-place  $z$ , then  $x$  in-same-place  $z$ .

Here, "any" corresponds to one of the two quantifiers that are used in the predicate calculus, the so-called "universal quantifier", which is sometimes symbolized as " $\forall$ ". The variables  $x$ ,  $y$ , and  $z$ , can have as values any individuals in the domain under discussion. Because this premise expresses a consequence of the meaning of "is in the same place

as", it is known as a *meaning postulate*. Other postulates can express matters of fact, e.g. "if x is a dog, then x needs a license".

The formal derivation of a conclusion in the predicate calculus depends on three stages:

1. eliminating the quantifiers from the premises;
2. reasoning with the **propositional** connectives;
3. re-introducing, if necessary, appropriate quantifiers.

Only one premise in our example (the meaning postulate) contains quantifiers, and they are all universal, i.e. "**any**". According to the rule of inference for eliminating a universal quantifier, if a predicate applies to any individual in the domain of discourse, then one can freely substitute the name of any specific individual in the domain in place of the quantified variable. Using this rule, we can replace the variables in the meaning postulate by a convenient choice of names:

∴ If Anna in-same-place Ben and Ben in-same-place Con then  
 Anna in-same-place Con.

We now proceed to the **propositional** stage of inference:

- ∴ Anna in-same-place Ben and Ben in-same-place Con.  
 (A conjunction of the two premises.)
- ∴ Anna in-same-place Con.  
 (Modus ponens from the previous two assertions.)

This conclusion is the one that we need: Anna is in the same **place** as Con, and so **the third stage—the restoration of quantifiers—is unnecessary.**

Many deductions depend on the other quantifier used in the predicate calculus, the so-called "existential quantifier" which logicians define as meaning, "at least some", and which they often symbolize as "∃". Thus, the premises:

Some Avon letters are in the same place as all Bury letters.  
 All Bury letters are in the same place as all Caton letters.

validly imply the conclusion:

Some Avon letters are in the same place as all Caton letters.

If a predicate applies to at least *some* individual in the domain under discussion, then there is a formal rule that permits the name of an individual to be substituted in pia

because the individual stands in for an existentially quantified variable, the name must not have occurred already in the argument, or else one may be led to a fallacious conclusion. Consider, for example, the premises:

Someone is tall.

(In **Loglish**: for at least some  $x$ ,  $x$  is a person and  $x$  is tall.)

Someone is not tall.

(For at least some  $x$ ,  $x$  is a person and  $x$  is not tall.)

"Anna" can be substituted for  $x$  in the first premise, but if the same name were then substituted in the second premise, there would be a contradiction: Anna cannot be both tall and not tall. In the third stage of deduction, when a quantifier is re-introduced in place of a name, one must use the rule that restores an existential quantifier for those names that have been introduced by the existential rule. Readers who wish to see a complete formal derivation in the predicate calculus should consult Tables 7.1 and 7.2 in Chapter 7.

The specification of a semantics for the predicate calculus is a more complicated business than the use of truth tables. It depends, as **Tarski** (1956) established, on characterizing *models*, which can concern **either** the real world or more abstract mathematical realms. The procedure is complicated because a model may contain an infinitude of different individuals (as in arithmetic where there are infinitely many **numbers**). For the sake of illustration, we will consider a finite model in which there are just three individuals: Arthur, Betty, and Carol, and one relation:  $x$  is in the same place as  $y$ , where  $x$  and  $y$  take individuals as their values. We stipulate that in our model:

Arthur is in the same place as Betty.

and, of course, that everyone is in the same place as themselves:

Arthur is in the same place as Arthur.

Betty is in the same place as Betty.

Carol is in the same place as Carol.

Logicians use two sorts of semantic rules for the interpretation of sentences. The first sort assign interpretations to basic terms, e.g. "**Arthur**" refers to the individual, Arthur, in the model; and "is in the same place as" refers to the set of pairs of individuals in the model who satisfy the relation, namely, Arthur and Betty, Arthur and Arthur, Betty and Betty; and Carol and Carol. The second sort of semantic rules work

in parallel to the **syntactic** rules defining the well-formed expressions in the calculus. These semantic rules build up the interpretation of a sentence in a way that depends on both the interpretation of its parts and the syntactic relations amongst those parts. Logicians refer to this system as a "compositional" semantics.

A key role is played by the semantic rules for quantified sentences. The syntax of a universally quantified sentence, such as:

For any  $x$ , Arthur is in the same place as  $x$

can be analyzed as having two constituents:

For any  $x$ ,  $S$

where  $S$  equals "Arthur is in the same place as  $x$ ".  $S$  is the *scope* of the quantifier and it binds any occurrence of  $x$  within its scope. Hence, the semantic rule for a universally quantified assertion states that the assertion is true if and only if replacing the occurrences of  $x$  in  $S$  by the name of *any* individual in the model results in a true sentence. Thus, the assertion is true provided that each of the following sentences is true in the model:

Arthur is in the same place as Arthur.

Arthur is in the same place as Betty.

Arthur is in the same place as Carol.

The first two sentences are true in our model, and the third is false. Hence, the quantified assertion is false.

There is an analogous semantic **rule** for existential quantification. An existentially quantified assertion:

For some  $x$ ,  $S$

is true if and only if replacing the occurrences of  $x$  in  $S$  by the name of *at least one* individual in the model results in a true sentence. Thus, the assertion:

For some  $x$ , Arthur is in the same place as  $x$

is true provided that at least *one* of the above triplet of sentences is true. In this case, the quantified assertion is true, because Arthur is in the same place as Betty (and himself).

The interpretation of assertions with two quantifiers, such as:

For some  $x$ , for any  $y$ ,  $x$  is in the same place as  $y$   
(i.e. **Someone** is in the same place as everyone.)

or:

For any  $y$ , for some  $x$ ,  $x$  is in the same place as  $y$   
(i.e. For everyone, someone in the same place as them.)

calls for a double application of the rules. At its highest level, the first of these sentences has the syntactic analysis:

For some  $x$ ,  $S$

and so it is true provided that  $S$  itself is true, i.e. provided that:

For any  $y$ ,  $x$  is in the same place as  $y$

is true. In short, we can interpret a sentence containing several quantifiers by, in effect, peeling them off one at a time, and looping through the substitutions within their respective scopes. When we get to the "bottom line", i.e. an assertion that does not contain variables, e.g.:

Arthur is in the same place as Betty.

its truth value is given by the basic **semantics**—the particular relations among the individuals in the model. The order of the quantifiers can obviously affect the interpretation of a sentence, because the quantifier with the larger scope is interpreted before the quantifier with the smaller scope. Hence, there is a difference in meaning between the two examples above.

Formal rules for the first-order predicate calculus, which we have now outlined, can be framed in a way that is complete, i.e. all valid deductions are derivable using them. This condition is not true, however, for the "second-order" predicate calculus in which properties can be quantified as well as individuals. This calculus is needed in order to give a full analysis of such assertions as, "Some sergeants have all the qualities of a great general", or of such unorthodox quantifiers as, "More than half" (see Barwise and Cooper, 1981). This logic is not complete: it cannot be formalized in a consistent way that guarantees that all valid deductions can be derived. Syntax is not equivalent to semantics.

One final point is vital before we can turn to the psychology of deduction. The alert reader will have noticed that we have described a semantic method for deduction in the **propositional** calculus (truth tables), but *not* one for deduction in the predicate calculus. Logicians **have** not proposed any deductive system that works directly with models of quantified sentences. A valid deduction must have a conclusion that is true in any possible model of the premises, and even a simple assertion about the real world, such as, "**The** cat sat on the mat", has infinitely many models, (Think of all the **different** possible configurations of cat and mat.) No **practical** procedure can examine infinitely many models in searching for a possible counterexample to a conclusion. Hence, what logicians have proposed are systems of *formal rules* based on the idea of such a search for counterexamples (see Beth, 1955; Hintikka, 1955; Smullyan, 1968). The method is simple, and has largely replaced "natural deduction" in textbooks. Each connective and quantifier has formal rules of inference that build up a search tree and that enable an avenue of exploration to be closed off whenever an inconsistency is encountered (see Jeffrey, 1981, for an excellent introduction, and Oppacher and Suen, 1985, for a computer implementation). But, the rules operate at one remove from models: they manipulate logical forms as do the rules of a natural deduction system.

What we aim to show in this monograph is: 1. that in everyday reasoning the search for counterexamples can be conducted directly by constructing alternative models; 2. that the psychological evidence implies that this procedure is used by human reasoners; and 3. that its simplicity and **capacity** to cope with certain finite domains make it an excellent method for the maintenance of systems representing knowledge in computers.

# 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 Romain, 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 demonstrate 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  
 $\therefore$  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.  $\therefore$  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  $x$ )  $\rightarrow$  ( (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  
 $\therefore$  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 Romain, 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<sub>1</sub>, P<sub>2</sub>, . . . P<sub>n</sub>**

Therefore, **P<sub>1</sub>** and **P<sub>2</sub>** and . . . **P<sub>n</sub>**.

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 Romain (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

	<i>Johnson-Laird</i>	<i>Braine</i>	<i>Rips</i>
<i>Conjunctions</i>			
A, B $\therefore$ A & B	*	+	+
A & B $\therefore$ A	+	*	*
<i>Disjunctions</i>			
A or B, not-A $\therefore$ B	+	*	+
A $\therefore$ A or B	+		*
<i>Conditionals</i>			
If A then B, A $\therefore$ B	+	+	+
If A or B then C, A $\therefore$ C		+	+
A $\vdash$ B $\therefore$ If A then B	+	*	+
<i>Negated conjunctions</i>			
not (A & B), A $\therefore$ not-B	+	+	
not (A & B) $\therefore$ not-A or not-B			+
A & not-B $\therefore$ not (A & B)	+		
<i>Double negations</i>			
not not-A $\therefore$ A	+	+	
<i>De Morgan's laws</i>			
A & (B or C) $\therefore$ (A & B) or (A & C)		+	
<i>Reductio ad absurdum</i>			
A $\vdash$ B & not-B $\therefore$ not-A	+	+	+
<i>Dilemmas</i>			
A or B, A $\vdash$ C, B $\vdash$ C $\therefore$ C		+	+
A or B, A $\vdash$ C, B $\vdash$ D $\therefore$ CorD		+	
<i>Introduction of tautologies</i>			
$\therefore$ A or not-A		+	+

## Notes

"+" indicates that a rule is postulated by the relevant theory.

"A  $\vdash$  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 experiment

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 (Johnson-Laird, 1989). Figure 2.1

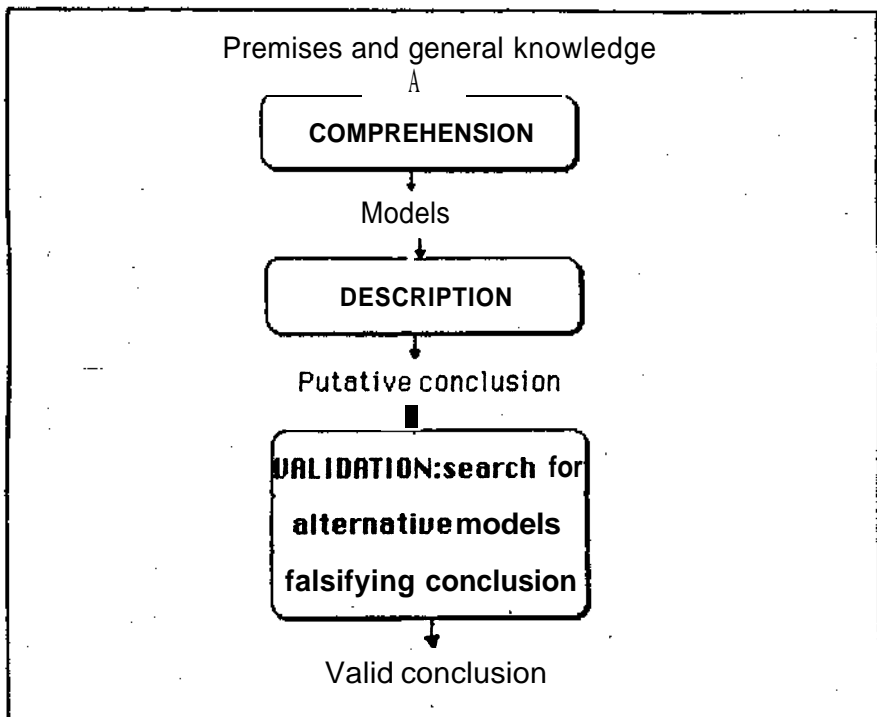


Figure 2.1. The three stages of deduction according to the model theory.

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 we

### The Euler circle representation of a syllogism

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

Each set of individuals is represented by a separate circle in the Euclidean plane

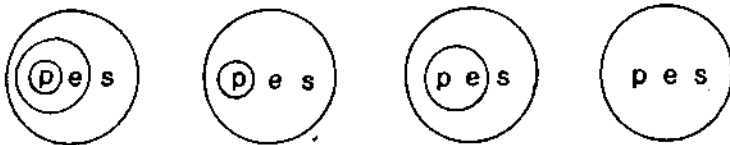
Premise 1 requires two diagrams



Premise 2 requires two diagrams



There are four ways of combining the sets of diagrams



It follows from all the combinations:

All psychologists are sceptics

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.

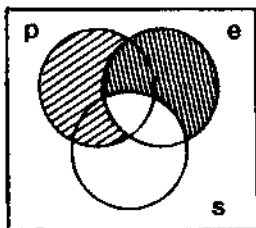
Deductions from singly-quantified premises, such as "All psychologists are experimenters", can be modelled using Euler circles (see Figure 2.2). Psychological theories have postulated such representations (Erickson, 1974) or equivalent strings of symbols

## 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.

# Reasoning with Propositions

Deductions based on propositional connectives, such as “not”, “if”, “and”, and “or”, are one of the main domains of human deductive competence, playing a part in many of the inferences in the other two main domains (relational and quantificational deductions). In this chapter, we will examine a number of clues from the psychological laboratory, and try to solve the mystery of how people reason **propositionally**. Because they can make deductions that do not depend on general knowledge, we can set aside theories based on content-specific rules. They are chiefly pertinent to the effects of content to be described in the next chapter. We are left with a choice between formal rules and mental models.

When people reason from conditionals, they are readily able to make a modus ponens deduction:

If there is a circle then there is a triangle.

There is a circle.

Therefore, there is a triangle.

They are less able to make the modus tollens deduction:

If there is a circle then there is a triangle.

There is not a triangle.

Therefore, there is not a circle.

Indeed, many intelligent individuals say that nothing follows in this case (see Wason and Johnson-Laird, 1972; Evans, 1982). The difference in difficulty between the two sorts of deduction is so robust that it demands an explanation. Rule theorists have two ways of explaining

phenomena: the choice of rules that they postulate as part of mental logic, and the relative availability or ease of use of these rules. In the present case, theorists assume that mental logic contains a rule for **modus ponens**:

**If A then B**

A

Therefore, B

but does not contain a rule for **modus tollens** (see e.g. the theories summarized in Table 2.2). In order to make the modus tollens deduction, it is therefore necessary to make a series of deductions. Given premises of the form:

If p then q

**not-q**

reasoners can hypothesise p:

P

(by hypothesis)

from which they can derive:

q

(by modus ponens from hypothesis and first premise)

This conclusion, together with the second premise, yields a self-contradiction:

**q** and **not-q**

(by conjunction)

The rule of *reductio ad absurdum* entitles reasoners to derive the negation of any hypothesis that leads to a self-contradiction:

not-p

(by *reductio*)

This chain of deductions is complicated, and so **modus tollens** should be harder than **modus ponens**.

The meaning of **propositional** connectives can be defined by truth tables, and, as we showed in Chapter 1, valid deductions can be made

by using the meanings of premises to eliminate contingencies from truth tables. But, logically-untutored individuals are unlikely to use this method. It calls for too many contingencies to be kept in mind, as theorists of all persuasions are agreed (Wason and Johnson-Laird, 1972; Osherson, 1975; **Braine** and **Rumain**, 1983). Indeed, people are notoriously bad at manipulating truth tables, they have difficulty in describing them, and they often fail to consider all possibilities in tasks analogous to the assessment of truth tables (Byrne and Johnson-Laird, 1990a). To abandon truth tables, however, is not necessarily to abandon the semantic approach to propositional reasoning. What is needed is a theory that reconciles the semantics of truth tables with the constraints of mental processing, and that does so in a way that explains human performance.

### MODELS FOR CONNECTIVES

Could there be a theory of propositional reasoning based on mental models? Because the approach was originally developed for spatial and **quantificational** reasoning, many critics have been skeptical (**Braine et al**, 1984; Rips, 1986; Evans, 1987). Happily, their skepticism has been overtaken by just such a theory. It assumes that people can form mental models of the states of **affairs** described in premises, but that they leave as much information as possible implicit in their models rather than spelling it out explicitly (**Johnson-Laird, Byrne, and Schaeken**, 1990). Given a conjunction describing what is on a blackboard, such as:

There is a circle and there is a **triangle**.

they build a single model of the **following** sort:

○    Δ

With a disjunctive premise, such as:

There is a **circle** or there is a **triangle**.

they **build** two alternative models to represent the possibilities:

○  
A

where we adopt the **notational** convention of putting separate models on separate lines. If, in addition to this disjunctive premise, someone asserts categorically:

**There isn't a circle.**

then reasoners can use this information to update the set of models. It eliminates the first model, which contains a circle. But it can be added to the second model:

$\neg O \quad \Delta$

where “ $\neg$ ” is a **propositional-like** tag representing negation. These tags may seem **odd**, particularly if one thinks of mental models as representing only physical or perceptible situations (Inder, 1987). In fact, **propositional** annotations are innocuous and easy to implement (see Polk and Newell, 1988; Newell, 1990), and they can be defended on psychological grounds (see Chapter 6).

The annotated model above corresponds to the description:

There is not a circle and there is a triangle.

One of the constraints on human deductive competence is parsimony, and so the procedure that formulates conclusions keeps track of categorical assertions and does not repeat them. It accordingly concludes:

There is a triangle.

which is valid because no other model of the premises falsifies it. Hence, a disjunctive deduction of the form:

p or q  
not-p  
Therefore, q

can be made by using the meanings of the premises to construct and to eliminate models. There is no need for formal rules of inference.

When psychologists test whether a disjunction, such as:

There is a circle or there is a **triangle**.

is interpreted inclusively or exclusively, they find that subjects do not respond in a uniform way. Typically, adults are biased towards an inclusive interpretation, but a sizeable minority prefer the exclusive interpretation (Evans and Newstead, 1990; Deane, 1970). The results

are not consistent from one experiment to another, though a semblance of consistency occurs if content (or context) suggests one or other of the two interpretations (see **Newstead** and **Griggs**, 1983). The lack of a consensus seems strange at first, as does the fact that people are normally aware neither of the two possible interpretations nor of settling on one of them as opposed to the other. The phenomena are still more puzzling viewed through the spectacles of rule theories, because these theories **presuppose** an initial recovery of the logical form of premises (see **Braine**, 1978; **Rips**, 1983), which makes explicit whether a disjunction is inclusive or exclusive.

The puzzle is resolved by the model theory. The initial representation of a disjunction by the models above is consistent with an inclusive or with an exclusive interpretation. The models can be fleshed out explicitly to represent either sort of disjunction. The distinction depends on making explicit that all instances of a particular contingency, e.g. those in which there are circles, have been exhaustively represented in the set of models. In other words, reasoners may know that there could be circles in other models, which they have yet to make explicit, or they may know that they have represented all circles explicitly. The contrast is a binary one, and we will use square brackets as our notation for the conceptual element corresponding to an exhaustive **representation**. Thus, the exclusive disjunction:

Either there is a circle or else there is a triangle, but not **both**.

has the following models:

[O]                    [Δ]

which represent explicitly all the contingencies containing circles and all the contingencies containing triangles. Hence, the further assertion:

There is a circle.

picks out the state of affairs in the first model. Because the triangles are exhausted in the second model, the first model can be "fleshed out" in only one way:

[O]    [¬Δ]

The procedure that formulates conclusions now yields:

There is not a triangle.

The explicit representation of the inclusive **disjunction**:

There is a circle or there is a **triangle**, or **both**.

allows for three alternative possibilities:

[O] [A]

[O]

[A]

including the joint contingency of circle and triangle. Where an item is not exhaustively **represented**—as shown in our notation by the absence of square brackets, then it is always possible to add further models containing that item. Hence, there is no need for the initial models of the disjunction:

○

△

to distinguish between inclusive and exclusive disjunction. One is at liberty to introduce or to exclude the joint contingency of circle and triangle.

A similar phenomenon occurs with conditionals. In a “binary” context, people interpret a conditional as implying its converse. **Legrenzi** (1970) demonstrated this point by using such conditionals as:

If the ball rolls to the left, then the red light comes on.

in a situation where the ball could roll either to the left or right, and the light was either red or green. But, when content and context are neutral, sometimes a conditional is taken to imply its converse, and sometimes not: people are neither consistent with one another nor from one occasion to another (see **Wason and Johnson-Laird**, 1972; **Staudenmayer**, 1975; **Staudenmayer and Bourne**, 1978; **Evans**, 1982). Again, it seems strange at first that there should be these vagaries in the interpretation of conditionals, and that one is not normally aware of them.

The model theory accounts for the phenomenon. A conditional, such as:

**If there is a circle then there is a triangle.**

calls for a model in which there is a circle (and thus a triangle), but the assertion is consistent with a state of affairs in which there isn't a circle. People do not initially make explicit the nature of this alternative, but merely represent its possibility in a second model that has no explicit content:



where the three dots denote a model with no explicit content. This second model allows for a subsequent explicit content, and it rules out a conjunctive description of the models. The categorical premise for modusponens:

There is a circle.

is accommodated within the set of models by eliminating the second model, because a circle occurs in the first model:



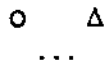
The model supports the conclusion:

There is a triangle.

which is valid because no other model of the premises falsifies it. If the categorical premise is instead one for modus tollens:

There is not a triangle.

then it can be accommodated within the initial models for the conditional:



only by eliminating the first model which contains a triangle. This process leaves only the second model, which now incorporates the information from the categorical premise:

$$\neg\Delta$$

Because this model represents only the categorical premise, it seems that nothing follows. This response is the most frequent error that subjects make given these premises. In fact, a valid deduction can be made but it is necessary, as we will now show, to flesh out the models before eliminating anything.

In the initial models of "If there is a circle then there is a **triangle**" neither the circle nor the triangle is exhaustively represented. But, the models can be fleshed out explicitly either as a conditional (i.e., material implication) or as a **bi-conditional** (i.e., material equivalence). The bi-conditional interpretation, "if and only if there is a circle then there is a **triangle**" calls for both the antecedent and the consequent to be exhausted:

[O]	[A]
-----	-----

...

The first **step** towards the conditional **interpretation** is to represent the antecedent as exhausted:

[o]	A
-----	---

...

These models can be further fleshed out to make explicit that a triangle can occur in the absence of a circle:

[O]	[A]
	[A]

...

The categorical premise:

There is not a triangle.

eliminates the first two models. Prior to this step, however, the models can be fleshed out completely:

[O]	[A]
[¬O]	[Δ]
[¬O]	[¬Δ]

Now the categorical premise eliminates the **first** two models and leaves behind only the third:

$[\neg O] \quad [\neg \Delta]$

This model supports the conclusion:

There is not a circle.

which is valid because no model of the premises falsifies it. The same procedure ensures that the same conclusion is drawn from the models of a **bi-conditional**. In both cases, *modus tollens* depends on fleshing out models and on detecting inconsistencies, and so it is a more complicated deduction than *modus ponens*.

Perhaps surprisingly, the difference in difficulty between *modus ponens* and *modus tollens* disappears when an implication is expressed by a statement using "only if" (see Evans, 1977b; Roberge, 1978), e.g.:

There is a circle only if there is a triangle.

This assertion has the same truth conditions as the conditional:

If there is a circle then there is a triangle.

because both are false only when there is a circle but no triangle. Evans and Beck (1981) suppose that the word "if directs a **reasoner's** attention to the proposition that follows it, irrespective of the occurrence of "only". Hence, forward inferences from antecedent to consequent are easier with conditionals, whereas backwards inferences from consequent to antecedent are easier with "only if" assertions. Braine (1978, p.6) offers an alternative account:

The behaviour of *p only if q* can be explained if we try to derive the meaning of *only if* as a compound of the meanings of *only* and *if*. In ordinary usage, *only* is equivalent to a double negative or **no ... other than** (e.g. *Only conservatives voted for Goldwater* = *No one other than conservatives voted for Goldwater*). We can use this equivalence to paraphrase *only* away from *p only if q*, for example, by the following steps: *p only if q* = **not p if other than q** = *if not q then not p*.

**One** trouble with both of these accounts is that they appear to predict a reversal: the difficulty of *modus ponens* and *modus tollens* rather than

its disappearance. The model theory, however, offers a straightforward explanation. Following **Braine's** intuition and a linguistic analysis of "only" (**Keenan**, 1971), the theory assumes that an assertion, such as "there is a circle only if there is a triangle", leads to two explicit models right from the start. One represents the positive contingency: if there is a circle then there is a triangle; the other represents the negative contingency: if there **isn't** a triangle then there **isn't** a circle:

[O]	A
$\neg O$	$[\neg \Delta]$
...	

These initial models allow both modus ponens and modus tollens to be made without any further fleshing out. Because two explicit models are required, both deductions should be more difficult than modus ponens with a conditional. The data confirm this prediction.

Table 3.1 summarizes the models for the different connectives, both the initial models and the completely explicit ones.

The same interpretations can be used to build up models of premises containing more than one connective. Here, for example, is one of the problems that **Braine et al.** (1984) asked their subjects to rate for difficulty:

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

Since there is no need for subjects to represent **exhaustiveness** for the problems used in this task, they are likely to have interpreted the first premise by building the following models:

C	P
H	P
...	

The second premise then eliminates all but the first model:

C	P
---	---

which supports the conclusion. This account is successful in explaining a number of aspects of the rating data that **Braine et al.'s** own theory leaves unexplained (see **Johnson-Laird, Byrne and Schaeken**, 1990).

**Table 3.1**  
 Models for the **propositional** connectives. Each line represents an alternative **model**, and the square brackets indicate that the set of contingencies has been exhaustively represented

1. <i>p and q</i>		
Initial model:	p	q
Explicit model:	[p]	[q]
2. <i>p or q</i>		
Initial models:	p	q
Explicit models:	Inclusive [p] [-q] [-p] [q] [p] [q]	Exclusive Cp] [-q] [-p] [q]
3. <i>If p then q</i>		
Initial models:	p	q
	...	
Explicit models:	Conditional [p] [q] [-p] [q] [-p] hq]	Bi-conditional [p] [q] [-p] [-q]
4. <i>p only if q</i>		
Initial models:	[p]	q
	¬p	[-q]
	...	
Explicit models:	Conditional [p] [q] [-p] hq] [-p] [q]	Bi-conditional [p] [q] hp] [-q]

The relation between models and truth tables should now be evident. Consider, for example, the truth table for an inclusive disjunction:

circle	triangle	circle or triangle, or both
T	T	T
T	F	T
F	T	T
F	F	F

An explicit set of models represents only those contingencies that are true:

[O]	[A]	-	the <b>first</b> line in the truth table
[O]	[¬A]	-	the second line in the truth table
[¬O]	[A]	-	the third line in the truth table

And with these contingencies only those elements that match the named constituents of the disjunction are represented at first:

[O]	[A]
[O]	
	[A]

These models are precisely the ones for inclusive disjunction. The essence of the theory is accordingly that people use models that make explicit as little information as possible, and in this way, they overcome the unwieldy bulk of truth tables.

The theory makes three processing assumptions. The first is that the greater the number of *explicit* models that a **reasoner** has to keep in mind, the harder the task will be: it will take longer, and will be more likely to lead to errors. The second assumption is that a deduction that can be made from the initial models of the premises will be easier than one that can be made only by fleshing out the models with explicit information. This process also takes time and places a load on working memory. The third assumption is that it takes time to detect inconsistencies between elements of models (see e.g. Wason, 1959; Clark and Clark, 1977).

## THREE PHENOMENA PREDICTED BY THE MODEL THEORY

### 1. Conditionals and Bi-conditionals

A **new** theory should lead to the discovery of new phenomena. The model theory does indeed make novel predictions, and so in this final part of the chapter we turn to them. We begin with a **difference** between conditionals and bi-conditionals. A conditional:

If Tony is in Kerry then Noel is in Dublin.

requires initially only one explicit model (and one implicit model), but when it is fleshed out to make a modus tollens deduction, it requires three explicit models. A **bi-condi**

If and only if Tony is in Kerry then Noel is in Dublin.

requires one explicit model for modus ponens, but only two for modus tollens. This **difference** leads to the prediction that modus tollens should be easier with a **bi-conditional** than with a conditional, but there should be no **difference** between the **two** for modus ponens, because both require only a single explicit model.

We tested this prediction in an experiment with sixteen adult subjects, who drew their **own** conclusions for two problems of each sort (see Johnson-Laird, Byrne, and Schaeken, 1990). The results confirmed the predictions. The percentages of correct conclusions were as follows:

Modus ponens with a conditional:	97%
Modus ponens with a bi-conditional:	97%
Modus tollens with a conditional:	38%
Modus tollens with a bi-conditional:	59%

Thus, modus tollens was easier with a bi-conditional than with a conditional, but there was no reliable difference between them for modus ponens.

Although a rule theory can accommodate these findings, it does so in an *ad hoc* way. The theory does not include a rule for modus tollens, and so the explanation cannot be based on the relative availability of such a rule for conditionals and **bi-conditionals**. Modus tollens depends on a chain of deductions. With our materials, the premises are of the form:

1. **If p then q**
2. **q'**, where q is incompatible with q', i.e.:
3. **If q then not-q'**

and the chain of deductions is as follows:

4. Suppose: p
5.  $\therefore q$   
(by modus ponens, from 1 and 4)
6.  $\therefore$  **not-q'**  
(by modus ponens, from 3 and 5)
7.  $\therefore$  **q'** and **not-q'**  
(by **conjunction**, of 2 and 6)
8.  $\therefore$  not-p  
(**reductio ad absurdum**, from 4 and 7)

Why should this sequence of inferential steps be easier with a **bi-conditional** than with a conditional? The experiment did not detect any difference between the two sorts of conditional for modus ponens (the **first** two steps in the derivation after hypothesizing *p*). The only feasible explanation seems to be that it is easier to think of making a hypothetical argument with a bi-conditional than with a conditional. A rule theory has no machinery to explain why this difference should occur.

## 2. Conditionals and Exclusive Disjunctions

The model theory predicts that it should be easier to argue from a conditional, such as:

If Linda is in Amsterdam then Cathy is in **Majorca**.

than to argue from an exclusive disjunction, such as:

Linda is in Amsterdam or Cathy is in **Majorca**, but not both.

The conditional calls for the initial construction of only one explicit model, whereas the disjunction calls for the initial construction of two explicit **models**—one representing Linda in Amsterdam, and the other representing Cathy in **Majorca**. The theory therefore predicts that, in general, deductions based on conditionals should be easier to make than those based on exclusive disjunctions, because disjunctions from the outset place a greater load on working memory. Some corroboratory evidence exists in the literature. Roberge (1978), for example, obtained such an effect, but his study was limited to only one sort of deduction. Evans and Newstead (1980) similarly report that when one constituent of a conditional is negated, **reasoners** can still cope, but when one constituent of a disjunction is negated they become hopelessly lost.

We have also tested the prediction that modus ponens should be easier than the analogous affirmative deduction based on an exclusive disjunction:

Linda is in Amsterdam or Cathy is in **Majorca**, but not both.

Linda is in Amsterdam.

What follows?

In addition, we tested the prediction that modus tollens should be easier than the analogous negative deduction based on an exclusive disjunction:

Either Steven is in Donegal or Jenny is in Princeton, but not both.

Jenny is in London.

What follows?

The model theory also predicts that these negative deductions should be harder than the affirmative deductions above, because the negative deductions call for the detection of an inconsistency between elements of models.

In principle, a negative deduction with a conditional calls for two or three models to be made explicit whereas with the disjunction it calls for only two explicit models, but the fleshing out of conditionals occurs after their *initial* interpretation, whereas reasoners should already have run into trouble with disjunctions. Will the two variables interact? The theory predicts **that** they should, because the **difference** between the two conditional deductions should be relatively large (one model versus **two** or three models) whereas the only difference between the disjunctive deductions is that the negative inference calls for detecting an inconsistency.

In our experiment, fourteen adults drew their own conclusions for four instances of each of the four sorts of problem (see Johnson-Laird, Byrne, and Schaeken, 1990). The results were clear. The percentages of correct conclusions were as follows:

Modus ponens:	91% correct.
Modus <b>to</b> llens:	64% correct.
Affirmative disjunction:	48% correct.
Negative disjunction:	30% correct.

As we had predicted, the conditional inferences were easier than the disjunctive inferences, and the affirmative inferences were easier than the negative inferences. Not a single subject violated either prediction. There was also a trend towards the predicted interaction though it did not reach significance.

Rule theorists can accommodate these observations by assuming that the rule for modus ponens is easier to use than the disjunctive rule. Such a hypothesis does not *explain* why the difference exists: it merely posits it, or provides a parameter that can be estimated from data and used to predict performance with other problems (e.g. Rips, 1983). In contrast, the model theory explains why one sort of deduction is easier than the **other**.

### 3. "Double Disjunctions"

A third group of phenomena is predicted by the model theory. If the number of alternative models to be kept in mind is large, then there should be a breakdown in deductive performance. In effect, the experimenter can overload working memory to the point where deduction ceases to be possible. The nature of the breakdown, however, should be revealing. The obvious way in which to increase alternative models is by introducing disjunctive premises; unlike conjunctions or conditionals, they immediately demand more than one explicit model.

**Wason** (1977) has devised a striking demonstration of the **difficulty** of keeping track of disjunctive alternatives. The subject is **presented** with four designs based on two shapes (diamond or circle) and two colours (black or white). The experimenter makes two assertions:

First, there is a particular shape and a particular colour, such that any of the four designs which has one, and only one of these features is called a THOG.

Second, the black diamond is a THOG.

The subjects' task is to classify as THOGs, or not THOGS, the three other shapes: the white circle, the black circle, and the white diamond. The experimenter's **meta-assertion** embraces four disjunctive possibilities:

1. If a design is a circle or else black, then it is a THOG.
2. If a design is a circle or else white, then it is a THOG.
3. If a design is a diamond or else white, then it is a THOG.
4. If a design is a diamond or else black, then it is a THOG.

The assertion that the black diamond is a THOG eliminates the fourth possibility because it refers to both properties, and it also eliminates the second one because it refers to neither property. Both of the remaining possibilities (1 and 3) yield the same classification: the white circle is a THOG, but neither the black circle nor the white diamond is a THOG. Few subjects solve the problem, and most decide that the white circle is *not* a THOG, and that the other two designs are either THOGs or of unknown status. The subjects fail to **envisage** the four disjunctive possibilities and attempt to make direct evaluations of the designs from the assertion about the black diamond. They appear to make the unwarranted inference:

If a black diamond is a THOG, then any design that is neither black nor a diamond is not a THOG.

In studies where the four disjunctive possibilities are spelt out explicitly for the subjects, their performance is reliably better (see Griggs and Newstead, 1982, but cf. Girotto and Legrenzi, 1989).

We have devised an informative demonstration that we call the "double disjunction" task, which the reader might like to attempt. It concerns the locations of three people, Linda, Mary, and Cathy:

Linda is in Cannes or Mary is in Tripoli, or both.

Mary is in Havana or Cathy is in Sofia, or both.

What, if anything, follows?

The most frequent response is that nothing follows (see Johnson-Laird, Byrne, and Schaeken, 1990). People indeed appear to be overwhelmed by the possibilities. When they do draw a conclusion it is seldom correct. In fact, the set of models for the first premise is:

[c]	[t]
[c]	
	[t]

where c denotes Linda in Cannes and t denotes Mary in Tripoli. The set of models for the second premise is:

[h]	[s]
[h]	
	[s]

where h denotes Mary in Havana, and s denotes Cathy in Sofia. At least one constituent proposition in each premise must be true, and so we can multiply out all the possibilities excluding only those cases where Mary is in Tripoli and Mary is in Havana, because one person cannot be in two places at the same time:

[c]	[t]	[s]
[c]	[h]	[s]
[c]	[h]	
[c]		[s]
	[t]	[s]

These models support the conclusion:

Linda is in Cannes or Cathy is in Sofia, or **both**.

or **equivalently**:

If Linda is not in Cannes then Cathy is in Sofia.

There is a simple way in which by giving subjects an *extra* premise the difficulty of a double disjunction is remarkably reduced. The phenomenon is predicted by the model theory. The reasoners are given a categorical premise such as:

Linda is not in Cannes.

in addition to the double disjunction above. They are then able to deduce that:

Mary is in **Tripoli**.

and that:

Cathy is in Sophia.

Why is the task so much easier? The answer is because it is no longer necessary to construct so many models. Given the models for Linda is in Cannes or Mary is in Tripoli:

[c] [t]  
[c] [t]

the categorical premise rules out the first two to leave only:

[t]

This model similarly eliminates all but one of the models for Mary is in Havana or Cathy is in Sofia.

[t] [s]

Hence the deduction can be made without having to construct the full set of five models. Of course, knowledgeable subjects could carry out a

double disjunction by spontaneously framing a hypothesis of their own in order to reduce the number of models they have to construct; the difficulty of the task implies that few logically-untrained individuals use this strategy.

When there is only one disjunction, deductions are more accurate from an exclusive disjunction than from an inclusive one. Newstead and Griggs (1983, p.97) argue that exclusive disjunctions are easier to grasp because the inferences are symmetrical: the truth of one constituent implies the falsity of the other, and *vice versa*. It is not clear, however, why this symmetry should make deduction easier. The model theory yields a simple alternative explanation: exclusive disjunctions call for a smaller number of explicit models than inclusive disjunctions. A double disjunction should therefore be reliably easier when the disjunctions are exclusive:

Linda is in Cannes or Mary is in Tripoli, but not both.  
 Mary is in Havana or Cathy is in Sophia, but not both.  
 What follows?

In this case, there are only three possible models:

	[t]	[s]
[c]		[s]
[c]	[h]	

But, the same conclusion as before is valid:

Linda is in Cannes or Cathy is in Sophia, or both.

We tested 24 adult subjects with both sorts of double disjunction (see Johnson-Laird, Byrne, and Schaeken, 1990). They drew 15% correct conclusions with the exclusive disjunctions, but only 4% with the inclusive disjunctions. The overwhelming majority of their conclusions suggested that they could imagine some of the possible models but not all of them. The following conclusion from a double inclusive disjunction, for instance:

If Mary is in Tripoli, then Cathy is in Sofia and Linda may be in Cannes.

is entirely accurate as far as it goes, but suggests that the subjects who drew it considered only two of the models. Similarly, the conclusion:

**Mary is in Tripoli and Cathy is in Sofia.**

is invalid, but it is consistent with three of the models. Figure 3.1 shows the percentages of conclusions consistent with one model of the premises, with two models of the premises, and so on. There are two striking features. First, the most frequent conclusions are consistent with just one model of the premises. Second, few conclusions are not consistent with any of the models of the premises. In short, the subjects do seem to be reasoning by constructing models (or at least one model) of the premises.

Once again, theories based on formal rules cannot account either for the difficulty of the task or for the sorts of errors that occur. The inclusive double disjunction has the form:

1.  $p$  or  $q$ , or both
2.  $q'$  or  $r$ , or both, where  $q'$  is incompatible with  $q$ , i.e.:
3. **If  $q$  then not  $q'$**

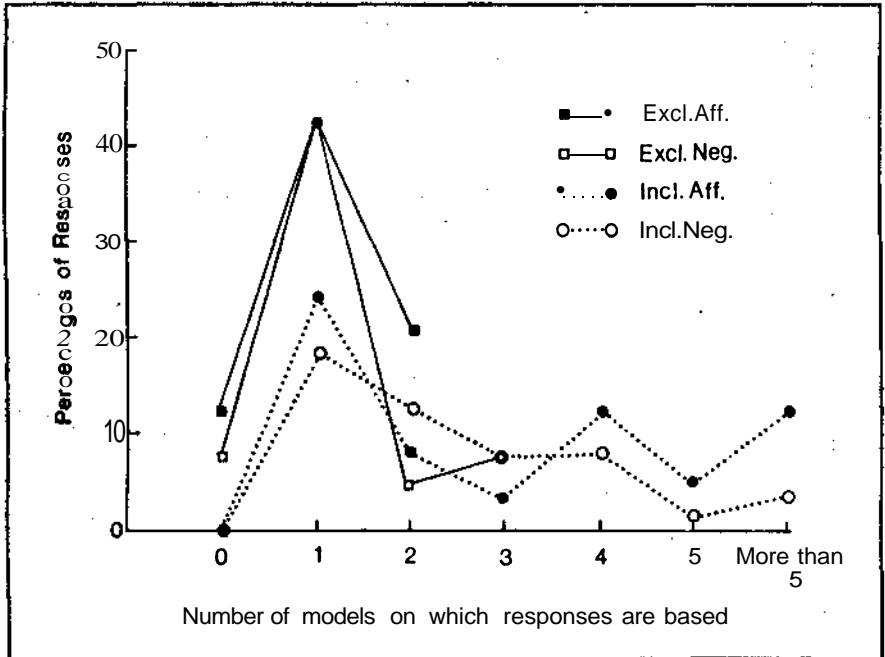


Figure 3.1. The percentages of responses in the double disjunction experiment that were consistent with one model of the **premises**, two **models**, and so on.

**A formal derivation proceeds as follows:**

4. Suppose: **not p**
5.  $\therefore q$   
(disjunctive rule, from 1 and 4)
6.  $\therefore$  **not q'**  
(modus ponens, from 3 and 5)
7.  $\therefore r$   
(disjunctive rule, from 2 and 6)
8.  $\therefore$  **If not-p then r**  
(conditional proof)

The derivation is no longer than the one above for modus **tollens**, and yet there is a massive difference in difficulty between the two deductions.

Perhaps the severest problem for theories based on formal rules is to account for the sorts of erroneous conclusions that are drawn. Psychiatrists **and-anthropologists** have sometimes proposed invalid rules of inference in order to explain irrational thinking (see e.g. **Levy-Bruhl**, 1910; von Domarus, 1944; and see also Jackendoff, 1988, for a similar speculation), but this tactic sacrifices deductive competence for too slight a gain in explanatory power. Errors with a "double disjunction" clearly cannot be explained in terms of misunderstanding the premises. Yet, if errors are supposed to be the result of a "spanner in the inferential works", then why do they correspond so reliably to a subset of the models of the premises? The answer appears to be that people try to reason by manipulating models but often succeed in **constructing** only one or two of the possible models of the premises.

## CONCLUSIONS

The model theory accounts for the existing phenomena of **propositional** reasoning and successfully predicts novel phenomena. It explains the following findings:

1. Disjunctions are interpreted in an indeterminate way if their content or context fails to elicit an inclusive or **exclusive** interpretation.
2. Conditionals are similarly interpreted in an indeterminate way.
3. Modus ponens is easier than modus tollens with conditional **premises**.

4. The **difference** disappears with “**only if**” premises, but both deductions are then slightly harder than modus ponens with a conditional premise.
5. Modus tollens is easier with a **bi-conditional** than with a **conditional**, but modus ponens is equally easy with both.
6. Deductions with conditionals are easier than deductions with exclusive **disjunctions**.
7. Double disjunctions are very difficult, but those with exclusive disjunctions are easier than those with inclusive disjunctions.

Different individuals reason in different ways. Some flesh out their models of a conditional in order to make a modus tollens deduction; others fail to do so. Some can construct several models of a “**double disjunction**”; others can construct only one model. The model theory predicts the general pattern of performance, and it could account for these individual differences in terms of such factors as the processing capacity of working memory. The theory has itself been modelled in two computer programs. The first implements our **psychological** assumptions: it uses implicit models until it is necessary to make them explicit, and it has a limited processing capacity. The second program, which we will describe in Chapter 9, is an exercise in artificial intelligence. It has no psychological constraints and so it draws a maximally parsimonious **conclusion from** any premises in the propositional calculus.

In this chapter, we have assumed an account of the meaning of connectives based on truth tables because our agenda has been dictated by theories based on formal rules. These theories are not intended to deal with *because* or *after* (or any connectives that lie outside truth tables). What is problematical is whether or not conditionals can be captured by truth tables. It is this and other puzzles of conditionals that **we now take up**.

## Conditionals

If there **hadn't** been guns I **wouldn't** have been **shot**.

If I hadn't been shot I wouldn't have been **re-elected**.

If I hadn't been re-elected, there wouldn't have been guns.

(Spitting Image's Ronald Reagan puppet)

*If*s problematical. Some connectives such as *and* and *or* have meanings that can be defined by truth tables, and apparent exceptions, such as the use of *and* to convey temporal sequence, can be explained as inferences based on general knowledge (Grice, 1975). Other connectives such as *because* and *when* cannot be defined by truth tables, because they invariably denote causal and temporal relations. But it is unclear whether *if* has a **truth-table** definition as we have provisionally assumed, or transcends such an analysis. The less that is known, the more that must be written; and the profundity of this puzzle is reflected in the **number** of books on conditionals (e.g. Harper, Stalnaker, and Pearce, 1981; Appiah, 1985; Traugott, ter Meulen, Reilly and Ferguson, 1986; Jackson, 1987). To explain reasoning with conditionals, however, we need to understand how they are understood. Our strategy will therefore be to outline a theory of the interpretation of "if", and then to show how this theory illuminates reasoning with conditionals. We begin with the meaning of neutral conditionals, which have no strong dependence on context or general knowledge for their interpretation, and then we consider other sorts of conditionals.

When people are asked to judge the truth or falsity of a neutral conditional, such as:

If there is a circle then there is a triangle.

they make the following characteristic judgements (see Johnson-Laird and Tagart, 1969; Evans, 1972):

<i>Situation</i>		<i>Judgement</i>
O	A	the conditional is true
O	*	the conditional is false
+	A	the conditional is irrelevant
+	*	the conditional is irrelevant

These judgements correspond to a "defective" truth table, and thus seem to bear out the view that a conditional makes an assertion conditional on the truth of its antecedent, and has no truth value when its antecedent is false (Quine, 1952; Wason, 1966). A related idea is that a conditional states a material implication, but a speaker who asserts a conditional knowing that its antecedent is false breaches the conventions of ordinary conversation (Grice, 1975; Lewis, 1976; Jackson, 1987).

Not all conditionals, however, have a defective truth table. A conditional with a negative **antecedent**:

If there isn't a circle then there is a triangle.

as even Stoic logicians noted, tends to be interpreted as equivalent to a disjunction:

There is a circle or there is a triangle.

A disjunction does have a complete truth table, and so these conditionals do too. A further class of conditionals lie outside truth tables altogether. They are **counterfactuals**, such as:

If there were a circle then there would be a triangle.

The truth or falsity of counterfactuals cannot depend merely on the truth or falsity of antecedent and consequent, because one normally takes for granted that these constituents are false, i.e. there **isn't** a circle, and there isn't a triangle. Although the linguistic distinction between indicative and **counterfactual** conditionals is not always clear cut (Dudman, 1988), theorists generally take counterfactuals to be true if, in any situation in which the antecedent is true, and which otherwise resembles the real world as closely as possible, the consequent is also true (see Stalnaker, 1968, 1981; Lewis, 1973; and Ripstein and Marcus, 1977).

In summary, some conditionals have a complete truth table, some have a defective truth table, and some seem to have no truth table at all. Not surprisingly, certain philosophers have abandoned the quest for their meaning. They do not have truth conditions, these skeptics claim, but only conditions in which it is justifiable to assert them, namely, provided that the consequent is highly probable given that the antecedent is true (see e.g. Adams, 1975; Appiah, 1985). Unfortunately for our purposes, if conditionals have no truth conditions then they cannot be true or false, and so they cannot occur in valid deductions, which by definition are truth preserving.

An ideal solution to the problem of "if" would be to establish a uniform semantics from which its chameleon-like interpretations would emerge. One distinguished proponent of formal rules, Braine (1979; Braine and O'Brien, 1989), has proposed such an account: a conditional is a rule of inference to the effect that its consequent can be inferred from its antecedent (see also Ryle, 1949, Chapter 5). The following cases, however, do not seem to warrant deductions from antecedent to consequent:

If you are interested then there is a Hitchcock movie on TV.

If you had needed some money then there was £10 in the desk.

but rather they are true or false merely in virtue of the truth values of their consequents. Similarly, one can judge the truth or falsity of a conditional, such as:

If there is a circle then there is a triangle.

by looking to see what is on the blackboard and without having to consider whether the consequent follows from the antecedent. Hence, we doubt the generality of Braine's thesis.

## THE MODEL THEORY OF THE MEANING OF CONDITIONALS

### Indicative Conditionals

The model theory provides an account of conditionals, but in order to establish it, we need to consider the metaphysics of everyday life, which distinguishes at least four sorts of situations: actual states of affairs, real possibilities, real impossibilities, and **counterfactual** situations. Actual **tates** of affairs, such as:

Mrs. Thatcher won the 1979 election.

are what happened. Real possibilities, such as:

Mrs. Thatcher wins the next election.

could happen—no matter how remote their likelihood—given the actual state of the world. Real impossibilities could never happen, given the actual state of the world:

Mrs. Thatcher won the American presidential election.

They must be distinguished from **counterfactual** situations, which were once real possibilities, but are so no longer because they did not occur:

Mrs. Thatcher lost the 1979 election.

The ability to envisage counterfactual situations is an important part of how one evaluates what actually happens (Hofstadter, 1985, p.239; Kahneman and Miller, 1986). A situation represented in a model can thus be treated by the interpretative system as actual, possible, impossible, or counterfactual. These categories, in turn, apply to both factual and fictional discourse. Granted their existence, we can set out our theory (which is a development from Johnson-Laird, 1986).

The interpretation of a **conditional's** consequent is identical to its interpretation as an isolated main clause in a context that satisfies the **conditional's** antecedent. It follows that antecedents must describe states of affairs, whereas consequents can have any **illocutionary force**—they can make assertions, ask questions, or give commands. The function of an antecedent is to establish a context, i.e. a state of affairs to be presupposed in interpreting the consequent. Hence, when one asks logically-naïve individuals to negate the conditional:

If there is a circle then there is a triangle.

their negation does not embrace the antecedent, and they assert instead:

If there is a circle then there is not a triangle.

The interpretation of an antecedent depends on its linguistic meaning and its **context**—in particular, the knowledge that is called to mind during the process of interpretation. An indicative antecedent, such as:

**If there is a circle ....**

calls for a model of a real possibility in relation to the current state of affairs, supplemented by an implicit model of an alternative, but real, possibility:

[O]

...

When these models have been established, the consequent can be interpreted in relation to the explicit antecedent model. Thus, the consequent:

...**there** is a triangle.

describes a state of affairs that obtains in the context described by the antecedent:

[O] A

...

This interpretation is consistent with subsequently fleshing out the models explicitly **as a conditional** or a **bi-conditional**.

The conditional is true provided that its consequent is true in any situation that satisfies the explicit antecedent model. The consequent is interpreted in a context where the antecedent is presupposed, and so the conditional has nothing to say—at least initially—about any alternative where the antecedent fails to hold, that is, the implicit model does not specify anything about the situation in which there is no circle. When the truth of a conditional is assessed in a situation where its antecedent is false, it is judged to be "irrelevant", because the models:

[O] A

...

do not specify anything explicitly about a situation in which there is no circle. This same lack of content, as we saw in the previous chapter, is responsible for the failure to make a **modus tollens** deduction.

Why should a negative antecedent, but not a negative consequent, lead to a disjunctive interpretation? A denial is plausibly used to correct a misconception, e.g. "A whale is not a **fish**" (see Wason, 1965), and so a negation is likely to call to mind the **affirmative** alternative.

Antecedents, unlike consequents, are exhaustively represented in models. Hence, a conditional with a negative antecedent:

If there **isn't** a circle then there is a triangle.

yields models in which the negative antecedent is **exhaustively** represented, and so models containing circles are the only alternative:

ho] A  
 [O]  
 [O]

This set of models is equivalent to the set for the disjunction: circle or triangle. Indeed, some subjects may merely represent the two positive instances:

A  
 [O]

A conditional with a negated consequent:

If there is a circle then there **isn't** a triangle.

only has models in which the negative consequent is exhausted when it is interpreted as a **bi-conditional**. Otherwise, it yields the models:

[O]  $\neg\Delta$   
 A  
 ...

which are not equivalent to the disjunction, circle or triangle, because the implicit model might contain neither a circle nor a triangle.

### Counterfactual Conditionals

The interpretation of **counterfactual** conditionals depends on invoking **counterfactual** situations. When someone asserts:

If there had been a circle ...

then listeners know that the speaker is assuming that there is not a circle, but that they are to **envisage** a **once possible**, but now counterfactual, context in which

Actual:	$\neg O$
Counterfactual:	[O]
	...

The consequent:

... there would have been a triangle.

implies that there is not a triangle, but asserts that in the Counterfactual situation where there is a circle, there is a triangle:

Actual:	• $\neg O$	$\neg \Delta$
Counterfactual:	[O]	$\Delta$
	...	

The implicit model allows that there was once a real possibility in which there **wasn't** a circle but there was a triangle. Speakers can overrule the **normal** interpretation of a Counterfactual (Wilson, 1975, p.122). The following assertion:

If there had been a circle then there would have been a triangle,  
and in fact there was a circle and so there was a triangle.

updates the model of actuality with the initial model of the Counterfactual set:

Actual:	O	$\Delta$
---------	---	----------

### The Relations Between Antecedent and Consequent

There is nothing too puzzling about the interpretation of neutral conditionals, because by definition there is no salient relation between antecedent and consequent. Yet no-one is likely to assert a conditional merely on the grounds that its consequent is true in the context specified by its antecedent. Despite the fact that it is true, one would hardly claim, for instance:

If Mrs Thatcher loses the election then one day it will rain.

On the **contrary**, one asserts a conditional where there are grounds for relating the antecedent to the consequent: the real possibility referred to by the antecedent has some bearing on the real possibility referred to by the consequent (cf. **Barwise**, 1989, Ch. 5).

The relations that can hold between antecedent and consequent may depend on a common referent or on general **knowledge** — a distinction that applies generally to relations in discourse (Johnson-Laird, 1983, Ch. 14). An example of a referential relation is:

If there is a circle then it is not large.

This relation rules out as impossible a certain contingency. There cannot be a case where the consequent of the conditional is false (the circle is large), and the antecedent is false (there **isn't** a circle). Hence, **modus tollens** is blocked. General knowledge enables certain rhetorical **effects** to be achieved by using conditionals with manifestly false constituents, **e.g.:**

If Thatcher wins the next election then pigs will fly.

Knowledge can also provide a framework that establishes a relation between antecedent and consequent (**Goodman**, 1947). The most important frameworks in which one state of affairs **constrains** another are those that establish inferential, causal, or deontic relations, **e.g.:**

If the number **hadn't** been divisible by 2 then it **wouldn't** have been even.

(An inferential relation.)

If the vase hadn't been dropped then it wouldn't have broken.

(A causal relation.)

If we hadn't promised then we needn't have gone.

(A deontic relation.)

In these cases, general knowledge informs the choice of what to represent in the models of the conditionals. We can illustrate this point by considering the difference between different sorts of causal assertion (Miller and Johnson-Laird, 1976, Sec. 6.3.5). On the one hand, the assertion:

If the vase hadn't been dropped then it wouldn't have **broken**.

makes a strong claim that one event caused another:

Actual:	d	b
Counterfactual:	$\neg d$	$\neg b$

The possibility of one event without the other is not even countenanced. On the other **hand**, an assertion of exactly the same syntactic form:

If the vase hadn't been fragile then it **wouldn't** have broken.

does not make a causal assertion, but rather stipulates that one state of affairs allowed another to occur:

Actual:	f	b
Counterfactual:	$\neg f$	$\neg b$
	f	$\neg b$

where the second Counterfactual model allows that even though the vase is fragile, it need not have broken, e.g., if it **hadn't** been dropped. The difference in the interpretation of these two conditionals depends, of course, on general knowledge. A third sort of conditional, such as:

If the vase hadn't been touched then it might not have **broken**.

makes only a weak causal claim:

Actual:	t	b
Counterfactual:	$\neg t$	$\neg b$
	$\neg t$	b

Causation is often treated by theorist and lay person alike as a primitive **unanalyzable** notion, but the concept can be taken to pieces in terms of real possibilities and Counterfactual situations (Miller and Johnson-Laird, 1976, Sec. 6.3.5). Table 4.1 summarizes the models for the main causal relations. Because models are related to truth tables, these analyzes could be expressed in a new form of truth table that distinguishes between real and Counterfactual contingencies. There is no need for any further decomposition of causation once models reflect the epistemological status of possibilities. "A caused B" in the strong sense means nothing other than that A happened and (then) B happened, and that had A not happened, then B would not have happened. In other words, it is a real impossibility for A to happen without B happening.

A similar analysis can be made of deontic matters. For example, the conditional:

Table 4.1  
The major causal relations. The actual and counterfactual models fall **within** the framework of physical principles

1. <i>Strong causation:</i>			
	a caused c	a prevented c	
Actual	a c	a $\neg$ c	
Counterfactual	$\neg$ a $\neg$ c	$\neg$ a c	
2. <i>Weak causation:</i>			
	a caused c	a prevented c	
Actual:	a c	a $\neg$ c	
Counterfactual:	$\neg$ a $\neg$ c	$\neg$ a c	
	$\neg$ a c	$\neg$ a $\neg$ c	
3. <i>Allowing relation:</i>			
	a allowed c	a allowed not c	
Actual:	a c	a $\neg$ c	
Counterfactual:	$\neg$ a $\neg$ c	$\neg$ a c	
	a $\neg$ c	a c	

If we had promised then we would have had to go.

makes a strong claim about obligation:

Actual:	$\neg$ p	$\neg$ g
Counterfactual:	p	g

The models in this case represent, not physical possibilities, but what is deontically possible, i.e., what is permissible. Physically it is, alas, all too possible to break promises, but as the models show, the only permissible course of action is to fulfil a promise. The relevant framework of **knowledge—physical**, deontic, or **inferential—effects** not only the interpretation of conditionals, but also the interpretation of modal auxiliary verbs such as “can” and “must” (Johnson-Laird, 1978). It leads to the fleshing out of different models of conditionals to reflect the underlying relation between antecedent and consequent.

The model theory of the meaning of conditionals can be summarized in three principles:

1. An indicative conditional is interpreted by constructing an **explicit** model of its antecedent, which is exhaustively represented, and to which is added a model of the consequent. An alternative implicit model allows for cases in which the antecedent does not hold.

2. A Counterfactual conditional is interpreted in the same way except that the models of its antecedent and consequent are of

counterfactual situations, and there is an explicit model of the actual situation.

3. Conditionals may elicit richer models in which more states are rendered explicit. This fleshing out of models occurs in several circumstances, e.g. when a referential relation, or one based on general knowledge, holds between antecedent and consequent.

According to this theory, a conditional is true if the proposition asserted by its consequent is true in the context described by its antecedent; it is false if the proposition asserted by the consequent is false in the context described by the antecedent. Knowledge, however, can establish that the antecedent condition is irrelevant to the truth or falsity of the consequent, e.g. "If you had needed some money then there was £10 in the desk". Where the antecedent of an indicative conditional is false, the interpretation will yield no truth value unless the models have been fleshed out to include explicit information about this possibility. The same principle applies *mutatis mutandis* to counterfactuals. They can have a complete truth table, but one contingency in it will be an actual state, one or more may be counterfactual situations, and at least one contingency will be a real impossibility.

## DEDUCTION WITH CONDITIONALS

### The Paradoxes of Implication

Our aim in formulating a semantics of conditionals was to elucidate deductions with them. Apart from the studies that we described in the previous chapter, the main phenomena of conditional reasoning are effects of content. There is, however, one formal problem that any theory of conditionals must confront. Material implication supports two valid deductions that appear to be paradoxical when the relation is expressed using a conditional. A false antecedent warrants a conditional with any consequent whatsoever:

1. You are not a millionaire.  
∴ If you are a millionaire then it will rain tomorrow.

Likewise, a true consequent warrants a conditional with any antecedent whatsoever:

2. The weather is fine.  
∴ If World War III started yesterday then the weather is fine.

If you doubt the validity of these **deductions**, then a glance at the truth table should settle your mind:

p	q	if p then q
T	T	T
T	F	<b>F</b>
F	T	<b>T</b>
F	F	T

As you see, the conditional is true whenever its antecedent is false (the last two rows of the table), and whenever its consequent is true (the first and third rows). Because either condition **suffices** to establish the truth of the conditional, the corresponding deductions **must** be valid. Theorists therefore face a **choice**: to abandon the truth-table analysis of conditionals (even if it is supplemented by inferences based on general knowledge), or to accept the validity of these apparently paradoxical **deductions** and to explain why they seem improper. We shall embrace the second alternative.

A conditional with a negated antecedent, as we have seen, has the same truth conditions as a disjunction. Hence, if someone asserts:

If Shakespeare didn't write the sonnets, then Bacon did.

and Shakespeare did write the sonnets, then the assertion is true. And, likewise, if Bacon wrote the sonnets, the assertion is true. Both of the following deductions are therefore valid:

1. Shakespeare wrote the sonnets.  
∴ If Shakespeare **didn't** write the sonnets then Bacon did.
2. Bacon wrote the sonnets.  
∴ If Shakespeare **didn't** write the sonnets then Bacon did.

They do not seem to be valid for the same reason that the corresponding deductions with disjunctions do not seem to be valid:

- 1'. Shakespeare wrote the sonnets.  
∴ Shakespeare wrote the sonnets or Bacon did.
- 2'. Bacon wrote the sonnets.  
∴ Shakespeare wrote the sonnets or Bacon did.

All four deductions throw semantic information away. They thus violate one of the fundamental constraints on human deductive competence (see

Chapter 2). And that is why they do not seem to be valid, even though they are.

### The Selection Task: Matching Bias

The best known phenomena of conditional reasoning occur in **Wason's** selection task (see e.g. Wason, 1966; Wason and Johnson-Laird, 1972; Wason, 1983). In the original version of the task, four cards are put in front of a subject. Each card has on its uppermost face a single symbol, say: A, B, 2, and 3; and the subjects know that every card has a letter on one side and a number on the other side. The experimenter then presents a neutral conditional:

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

The **subjects'** task is to select those cards that they need to turn over in order to determine whether the rule is true or false.

The majority of subjects select the A card, or the A and the 2 cards. Surprisingly, they fail to select the card corresponding to the case where the consequent is false: the 3 card. Yet, if the A had a 3 on its other side, the rule would be false; and so, by parity of reasoning, if the 3 had an A on its other side, the rule would also be false. In short, the correct selection consists in the card that renders the antecedent true and the card that renders the consequent false, because the combination of true antecedent and false consequent shows that the conditional itself is false.

The selection task calls for more than a deduction: subjects have to explore different possibilities, to deduce their consequences for the truth falsity of the rule, and on this basis to determine which cards to select.

This task has generated a large literature, which is not easy to integrate, and one **major** investigator, Evans (1989), has even wondered whether the paradigm tells us anything about deduction as opposed to heuristic biases. Part of **Evans's** concern arises from a phenomenon that he and his colleagues discovered, the so-called "**matching**" bias. When conditionals contain negated antecedents or consequents, subjects appear to ignore the presence of the negation in constructing an instance that falsifies the conditional. They merely match their selections to the cards mentioned in the conditional. They do not ignore negations, however, in constructing instances to render a rule true, or in coping with other connectives, such as disjunction.

If people succumb to matching bias in constructing false instances of conditionals, then they are also likely to succumb in carrying out the

selection task. Evans and Lynch (1973) confirmed this prediction. Given a **rule**, such as:

If there is not an S on one side of a card then there is a 9 on the other side.

then about two thirds of the subjects select the S card, which renders the antecedent proposition false. These subjects thus appear to ignore the negation. A comparable effect occurs with the rule:

If there is an S on one side of a card then there is not a 9 on the other side.

Most subjects select the card that renders the antecedent true, and over half of them select the card that renders the consequent false. In this case, the subjects are correct, though presumably because they are merely matching their selections to the items mentioned in the rule.

Evans (1989, p.33) argues that the matching bias is "a complex, linguistically determined relevance judgement rather than a simply (sic) availability or response priming effect." He explains it by making two hypotheses. First, a conditional statement:

If (not) p then (not) q

is always about p and q regardless of the presence of negation. This assumption is similar to our earlier argument that a negation leads to a representation of the affirmative case too. Second, the effect is overridden in a verification task by another, more powerful, linguistic factor: "The use of *if* invites one to entertain the supposition that the antecedent condition is **true** ... the listener is strongly invited to consider the hypothesis (mental model, possible world) in which the antecedent and consequent conditions are actually fulfilled." (Evans, 1989, p.32).

The matching bias, as Evans allows, can be reconciled with the model theory. We argued earlier that conditionals with negative constituents elicit representations of the corresponding positive items. Hence, a conditional with a negative antecedent:

If there is not an A then there is a 2

may elicit the models:

Similarly, a conditional with a negated consequent:

If there is an A then there is not a 2

may elicit the models:

[A]  
 2  
 ...

As we will see, these interpretations then yield the observed phenomena. The subjects are reasoning, but their representation of the conditional has been influenced by the presence of negation.

### The Selection Task: Realistic Content

With certain realistic rules or regulations, such as:

If a person is drinking beer then the person must be over 18.

subjects are more likely to reason correctly in the selection task. They choose the card corresponding to the case where the consequent is false: not over 18 (Griggs and Cox, 1982). Theories based on formal rules, as Manktelow and Over (1987) argue, cannot easily account either for the failure to select the false consequent with a neutral conditional or for its selection with these realistic conditionals. There is no difference in the logical form of realistic and neutral conditionals that could account for the results. Moreover, those arch-formalists, the Piagetians, claim that children have a capacity for falsification as soon as they attain the level of formal operations. Piaget describes this ability in the following terms: to check the truth of a conditional, if  $p$  then  $q$ , a child will look to see whether or not there is a counter-example,  $p$  and not- $q$  (Beth and Piaget, 1966, p.181). Yet adults conspicuously fail to do so with neutral conditionals in the selection task.

Several reasons have been put forward to explain why a realistic conditional may elicit the correct selection. They are all variants on the theory that people use content-specific rules of inference. The simplest hypothesis is that subjects recall specific counterexamples from their memory of actual or similar cases, and use this knowledge to guide their selection (Griggs and Cox, 1982). A more plausible hypothesis allows for analogies (e.g., D'Andrade, cited in Rumelhart and Norman, 1981; Manktelow and Evans, 1979; Griggs, 1983; cf. also Riesbeck and Schank, 1989). Although both hypotheses lack clear boundary conditions, neither

of them seems to embrace the **finding** that a general deontic framework can also improve performance. This **finding** was predicted by a third variant on content-specific rules: "pragmatic reasoning **schemas**" (Cheng and **Holyoak**, 1985; Cheng, **Holyoak**, Nisbett, and Olivier, 1986). These are rules of inference supposedly induced from experience, and they concern causation, permission, and obligation. The permission schema, for example, consists in four rules (or productions in a production system):

1. If the action is to be taken then the precondition must be satisfied.
2. If the action is not to be taken then the precondition need not be satisfied.
3. If the precondition is satisfied then the action may be taken.
4. If the precondition is not satisfied then the action must not be taken.

When a conditional, such as:

If a person is drinking beer then the person must be over 18.

cues the schema, then rule 4 directly elicits the idea that if a person is not over 18 then they must not drink beer. And this rule leads to the selection of the card corresponding to the false consequent: not over 18.

A variant of this idea has been advanced by Cosmides (1989), who argues that human evolution has led to a specific inferential module concerned with violations of social contracts. She shows that a background story eliciting such ideas can lead subjects to make a surprising selection: they choose instances corresponding to not-p and q for a conditional rule of the form, if p then q. In the context of the story, the rule:

If a man has a tattoo on his face then he eats cassava root.

tends to elicit selections of the following cards: no tattoo, and eats cassava root. We believe that there is a simple alternative explanation for this result. The subjects treat the rule as meaning:

A man may eat cassava root only if he has a tattoo.

Such an assertion, as we argued in the previous chapter, calls for models of the following sort:

[eating cassava]	tattoo
$\neg$ eating cassava	[ $\neg$ tattoo]

...

which, as the next section shows, leads to the **subjects'** choice of instances. There is no need to postulate a specific inferential module concerning the violation of social contracts (see also Griggs, 1984; Cheng and Holyoak, 1989).

Knowledge undoubtedly influences deduction, but is it represented by content-specific rules? There is no evidence for this form of representation; it could be represented by general assertions, which are used to construct models of the sort shown in Table 4.1. One further difficulty with pragmatic schemas is their use of **unanalyzed** modal auxiliaries, such as "may and "must", which could be analyzed in terms of possible and permissible states of affairs.

### The Model Theory of the Selection Task

The model theory assumes that reasoners use their knowledge, however it is represented, in constructing models of premises. The selection task is carried out in the following way:

1. The subjects consider only those cards that are explicitly represented in their models of the rule.
2. They then select those cards for which the hidden value could have a bearing on the truth or falsity of the rule.

This account is a simple modification of an earlier theory (Johnson-Laird and Wason, 1970) in which models of the rule now serve in place of truth tables.

A critical factor according to this theory is whether the models include explicit representations of negative instances. A neutral conditional, such as:

If there is an A then there is a 2

yields the models:

[A] 2

...

The subjects will consider both cards, but will select only the "A" card, because it alone has a hidden value that could bear on the truth or falsity of the conditional. If they interpret the rule as a **bi-conditional**:

[A] [2]

...

then they will select both cards. Rules with a negated antecedent or a negated consequent, as we argued earlier, tend to elicit models of the positive items, e.g:

2

[A]

and so subjects will tend to make the same selections as those for affirmative conditionals. Hence, performance with neutral conditionals reflects a bias towards selecting those cards that match the cards in the explicit model (**cf.** Evans, 1987; 1989; **Klayman** and Ha, 1987). There may be an independent bias towards verifying the rule (Wason and Johnson-Laird, 1972). And, according to the model theory, the card corresponding to the false consequent will be selected only if the models of the conditional are fleshed out to represent it explicitly, e.g:

[A] 2  
 -2

where since [A] is exhausted in the first model  $\neg 2$  must occur with  $\neg A$ . An insightful performance may further depend on an explicit representation of what is not possible, i.e. the real impossibility given the rule:

[A] [ $\neg 2$ ]

In short, the model theory predicts that people will select the card falsifying the consequent whenever the models are fleshed out with explicit representations of that card. This prediction makes sense of the **five** experimental manipulations that have been found to yield the correct selections, and neither memory for counterexamples nor pragmatic reasoning schemas can account for all of them:

1. Change the form of the rule. This manipulation includes the use of rules, such as "All circles are **black**", that elicit models of a single entity rather than of two separate entities (Wason and Green, 1984). It also

includes the use of conditionals with negated consequents, or the use of disjunctions (Wason and Johnson-Laird, 1969). We predict that changing the rule to an "only if" formulation should also enhance performance.

2. Change the content of the rule. This tactic includes the use of contents that trigger memories for violations (Johnson-Laird, Legrenzi, and Legrenzi, 1972), or memories for analogous events (Griggs and Cox, 1982; Klaczynski, Gelfand, and Reese, 1989).

3. Change the context of the rule. This tactic includes the use of a relevant deontic framework for the interpretation of the rule (Cheng and Holyoak, 1985; Cheng et al, 1986; Cosmides, 1989). General knowledge need not be represented by pragmatic reasoning schemas; it can nevertheless lead to the explicit representation of negative instances (see Table 4.1).

4. Change the content of the cards, by labelling them explicitly with negations (Cheng and Holyoak, 1985, Expt 2; Jackson and Griggs, 1990).

5. Change the task so that subjects are more likely to envisage all the alternatives explicitly. This manipulation includes reducing the choice to one between the consequent and the negated consequent (Johnson-Laird and Wason, 1970; Wason and Green, 1984; Oakhill and Johnson-Laird, 1985a), instructions to test violations of the rule (Valentine, 1985; Chrostowski and Griggs, 1985), and the verbalization of the reasoning behind one's selections (Berry, 1983; but cf. Klaczynski et al, 1989).

### The Suppression of Valid Deductions

The selection task challenges formal theories because the only effects of content that they can explain are those on the interpretation of premises. A more recently discovered effect of content challenges the foundation of all formal theories: the assumption that the rule of modus ponens is part of mental logic. To describe the effect, we need first to consider certain fallacies.

Ordinary reasoners given the premises:

If she meets her friend then she will go to a play.  
She did not meet her friend.

tend to draw the conclusion:

Therefore, she did not go to a play.

The inference is known (after its categorical **premise**) as denying the **antecedent**, and it is fallacious unless the first premise is interpreted as a **bi-conditional**. The following premises:

If she meets her friend then she will go to a play.

She went to a play.

lead to an analogous fallacy known (after its categorical premise) as affirming the consequent:

Therefore, she met her friend.

These fallacies might **have** led rule theorists to suppose that people are equipped with two invalid rules of inference for material implication:

1. If  $p$  then  $q$   
 $\text{not-}p$   
 $\therefore \text{not-}q$

and:

2. If  $p$  then  $q$   
 $q$   
 $\therefore p$

Rule theorists, however, have not in general adopted this idea, because they have been able to suppress the fallacies. They can do so by presenting an extra premise that establishes an alternative antecedent bringing about the same consequent (**Rumain, Connell, and Braine, 1983; Markovits, 1984, 1985**). Thus, where the original conditional is:

If she meets her friend then she will go to a **play**.

the additional presentation of:

If she meets her brother then she will go to a play.

blocks the bi-conditional interpretation of the original conditional. The subjects realize that she could have gone to a play even if she did not meet her friend. When the two conditionals are accompanied with the

appropriate categorical premise, subjects tend no longer to deny the antecedent or to affirm the consequent.

If the fallacies can be blocked by providing extra information, then, as rule theorists such as Romain, Connell, and Braine (1983) have argued, there cannot be mental inference rules corresponding to them. But, suppose that additional information could suppress valid deductions. What, then? By their own argument, rule theorists ought to claim that there cannot be inference rules for them, either. The question arises because Byrne (1986, 1989a) has devised a simple manipulation that suppresses modus ponens and modus tollens.

Given the conditional:

If she meets her friend then she will go to a play.

and the appropriate categorical premise, nearly everyone makes the modus ponens deduction, and a substantial proportion of people make the modus tollens deduction. Byrne suppressed these valid deductions by presenting an extra premise with the original conditional:

If she meets her friend then she will go to a play.

If she has enough money then she will go to a play.

The new premise reminds subjects of an additional condition that is necessary to bring about the consequent. Thus, both antecedents seem jointly necessary for the consequent to occur. When subjects are presented with these two conditionals and the categorical premise, they tend not to make the modus ponens deduction. The group of subjects who received only the original conditional tended to make modus ponens (96%), but the group who received the additional premise showed a striking suppression of the deduction (38%). There was a similar suppression of modus tollens. As one would expect, however, the new premise did not suppress the fallacies.

A distinguished rule theorist describes his concept of a formal rule of inference as follows:

By a formal logical rule, I take it, we mean a rule that applies to a string in virtue of its form. That is, the rule can apply whenever a string is described as having a certain **form**.... The question of whether there is a psychological version of this rule in the minds of normal people (not trained in logic) turns on whether they have a secure intuition, applying equally to any content, that [the rule applies]. I take it that they have. And for me, that's an end of it.

Byrne has shown that people do not have a secure intuition that modus ponens applies equally to any content. Hence, **by Macnamara's** criterion, one may doubt the existence of a rule for modus ponens in mental logic.

The suppression of modus ponens casts no doubt whatsoever on its validity as a rule of inference, but it does support our thesis that people make deductions not by following such rules **but** by building models. For the conditionals that suppress the fallacies, their knowledge leads them to models in which each alternative antecedent brings about the same consequent. For the conditionals that suppress the valid deductions, the **subjects'** knowledge leads them to construct one model in which both antecedents occur and an implicit alternative model. **These** premises therefore suppress the valid deductions because only one of the necessary conditions is asserted categorically.

#### The Spontaneous Use of Conditional Descriptions

One final question about the various theories: what do they imply about the spontaneous use of conditionals in descriptions? For rule theories, people should use an assertion of the **form**, if p then q, whenever q can be inferred from p (**Braine, 1979; Braine and O'Brien, 1989**). We have found that when subjects are asked to summarize a truth table succinctly, they hardly ever use a conditional to describe material implication or equivalence. According to the model theory, a conditional has the initial models:

[p]     q

...

and so a truth table presents too much information to assimilate, and also too many contingencies that have no explicit representation in the models. What should elicit a conditional, however, is a set of contingencies that corresponds to these two models. But, how is one to convey the content of the implicit alternative model?

Over a series of experiments, we developed a procedure in which the subjects have to paraphrase sentences (Byrne and Johnson-Laird, 1990a; Byrne and Johnson-Laird, 1990b). In the critical experiment, we presented subjects with sets of three simple sentences, such as:

- Laura has an essay to write.     (e)
- The library stays open.     (l)
- Laura studies late in the **library**.     (s)

and their task was to combine them into a single sentence using any words whatsoever. Three factual sentences, such as these, are likely to yield a model of an actual sequence of events:

e l s

and this model should be described using factual conjunctions, such as "and", and "when". To elicit a conditional, we needed to suggest an implicit alternative, and so we used a modal verb that signified a real possibility. In place of the simple factual assertion:

Laura studies late in the library.

we used the modal assertion:

Laura can study late in the library.

which implies that studying late is a possibility rather than a fact. We tested a group of 9 adult subjects with the factual sentences and a separate group of 18 subjects with the modal sentences. The group with factual materials used conditional descriptions on only 2% of trials, but the modal group used them on 36% of trials.

## CONCLUSIONS

Conditionals are problematical, but the theory of mental models makes sense of them. It accounts for how people understand them, how they reason with them, and how they use them in describing the world. Their initial interpretations, especially with a neutral conditional, produce one explicit model of the antecedent and consequent, and one implicit model that allows for alternative possibilities. These models lead to a defective truth table, an inability to make a modus tollens deduction, and a lack of insight into the selection task. When the models are fleshed out with explicit information, particularly from a knowledge of the relations between events, then judgements conform to a complete truth table, modus tollens is deduced, and an insightful choice in the selection task become feasible.

# Reasoning about Relations

Comparisons are invidious, but they are the **stuff** of many deductions in daily life:

Harold was better than Ted.  
Maggie was worse than Ted.  
Therefore, Harold was better than Maggie.

They hinge on relations that are *internal to propositions*, and so the **propositional** calculus cannot **capture their validity, because** it is not sensitive to the internal structure of premises (see Chapter 1). A proper logical analysis calls for the resources of the predicate calculus. Nevertheless, such **relational** deductions are easy to make, and seldom elicit errors. They are trivial, although it has proved far from trivial to understand how people make them.

Our plan in this chapter is to begin with these simple relational deductions, so-called "three-term series" problems. We will discover that despite many experimental studies these problems are too impoverished to discriminate among competing accounts of how people solve them. And so we will turn to more complex relational deductions that depend on at least two dimensions, e.g.:

The volume control is on the right of the tone control.  
The tuner is above the tone control.  
The clock is above the volume control.  
Therefore, the tuner is on the left of the clock.

We will report empirical studies of these deductions that have enabled us to reach a conclusion about the reasoning mechanism.

### THREE-TERM SERIES PROBLEMS

When logicians analyze relations, they focus on their logical properties. For example, the relation "in front of" is transitive, that is, it supports valid deductions of the form:

A is in front of B.  
 B is in front of C.  
 Therefore, A is in front of **C**.

Intransitive relations, such as "directly on top of", also support valid deductions:

A is directly on top of B.  
 B is directly on top of C.  
 Therefore, **A** is *not* directly on top of **C**.

Non-transitive relations, such as "next to", do not warrant either sort of deduction. If:

**A is next to B.**  
 B is next to C.

then A may be next to C if they are arranged in a circle, or it may not be if they are arranged in a line. Hence, there is no valid conclusion.

Other logical properties include symmetry and **reflexivity**, and their cognates. Symmetric relations, such as "in the same place as", give rise to the following sort of deduction:

A is in the same place as B.  
 Therefore, B is in the same place as A.

Reflexive relations, such as "identical to", yield the following sort of logical truth for any individual:

A is identical to A.

Relations yield still other sorts of deduction, and many of them have no recognized logical label, e.g.:

The police managed to prevent the man from assassinating the prime minister.  
 Therefore, **the** police prevented the man from assassinating **the** prime minister.  
 Therefore, the man did not assassinate the prime minister.  
**Therefore**, the man did not kill the prime minister.

We have outlined three broad approaches to the psychology of reasoning: **formal rules**, content-specific rules, and **mental** models. Theories of **relational** reasoning also fall **into** these main categories. Rule theories require representations of the logical properties of relations, which can be expressed, as we saw in Chapter 1, by meaning postulates. **Meaning** postulates for the relation, "in front of, for example, **would include** the following in which each variable is universally quantified:

- If x is in front of y and y is in front of z then x is in front of z.  
 (transitivity)  
 If x is in front of y then y is not in front of x.  
 (asymmetry)  
 If x is in front of y then y is behind x.  
 (converse **relation**)

Alternatively, there could be just one general postulate for each logical property, e.g.:

**If  $xRy$ , and  $yRz$ , then  $xRz$**   
 (transitivity)

and all relevant relations could be tagged to indicate which postulates **apply to them** (Bar-Hillel, 1967).

A deduction, such as:

1. The circle is in front of the triangle.
  2. The cross is behind the triangle.
- Therefore, the circle is in front of the cross.

can be made by first instantiating the meaning postulate that allows a relation to be transformed into its converse:

3. If the cross is behind the triangle, then the triangle is in **front** of the cross.  
 (instantiation of conversion postulate)

and by instantiating the postulate for transitivity:

4. If the circle is in front of the triangle and the triangle is in front of the **cross**, then the circle is in front of the **cross**.  
(instantiation of transitivity postulate)

The conclusion can then be proved using the standard rules for connectives:

5. The triangle is in front of the **cross**.  
(modus ponens from 2 and 3)
6. The circle is in front of the triangle and the triangle is in front of the **cross**.  
(conjunction of 1 and 5)
7. The circle is in front of the **cross**.  
(modus ponens from 4 and 6)

Where rule theories diverge is in **their representation** of logical properties, such as transitivity. They could be captured in postulates, which are assertions to be taken as true, or in content-specific rules of inference, which enable one assertion to be derived from others. The latter often take the form of productions in computer programs (see Chapter 2), **e.g.:**

(Condition (And (in front of x **y**)(in front of y z))  
(Action (in front of x z))).

Such programs have indeed been proposed for spatial inference (e.g., Ohlsson, 1981, 1984; Hagert, 1983, 1984; Olson and Bialystok, 1983).

Historically, the first theory of how people could carry out a three-term deduction was proposed by Hunter (1957). His starting point was William **James's** idea (1890, p. 646) about a series of the form:

$$a > b > c \dots > z$$

James argued that "any number of intermediaries may be expunged without obliging us to alter anything in what remains **written**". Hunter's theory adopts the same principle, but posits two operations to bring the premises into the required linear order: the conversion of a premise from the form,  $b < a$ , to the form,  $a > b$ , and the re-ordering of premises from, **say:**

$b > c$

$a > b$

into the order:

$a > b$

$b > c$

The rule for conversion depends on content because not all premises can be **validly** converted; and the expunging of intermediaries is similarly content-specific. Hence, Hunter's operations are akin to content-specific rules.

Clark (1969) proposed an explicitly content-specific theory. It stresses the factors that lead to greater difficulties in relational deductions, as reflected in longer response times. In certain **antonymic** pairs of expressions, such as:

**better - worse**

**taller - shorter**

**bigger - smaller**

the first term can be used in a neutral way to refer to the relative positions of two items on the dimension in question, whereas the second term conveys information about which end of the dimension the pair is to be found. To assert, for example:

Maggie is worse than **Ted**.

suggests that neither of them is much good. **Psycholinguistic** evidence has shown that these so-called "marked" terms are slightly harder to understand than their related "unmarked" terms. And Clark demonstrated the same phenomenon in three-term series problems. He also hypothesized that the specifics of a relation are harder to comprehend than the general information it conveys about the dimension. For example, one readily grasps from the previous assertion that Maggie and Ted are bad, but it is a little harder to determine their relative demerits. From these principles, he was able to predict the difficulty of different **three-term** series problems. His theory, however, concerns factors affecting performance, and was not intended to account for the complete sequence of processes that lead to the right answer.

Model theories of relational deductions were also among the earliest to be proposed (e.g. **DeSoto**, London, Handel, 1965; **Huttenlocher**, 1968).

We have implemented a computer program that is based on the model **theory** (an extension of one described in Johnson-Laird, 1983, Ch. 11), and that can carry out three-term and other, more complicated, relational deductions. It uses neither meaning postulates nor rules of inference. The transitivity of the relation, "in front of, for example, is nowhere explicitly represented, but is an emergent property from the meaning of the relation and its use in constructing models. The meaning is itself represented by a piece of code that is used by the compositional procedures that combine meanings according to **syntax** (see Chapter 9). The meaning of the premise is used to build models, to verify assertions in models, and to search for alternative models that falsify putative conclusions. It **states**, in effect, the direction in which a spatial model should be scanned in order to establish that one entity is in front of another from the **observer's** point of view, i.e. the positions on the line of sight should be scanned while holding the horizontal axis constant. Expressions such as "in front of have, in addition to this *deictic* meaning that depends on the speaker's point of view, a meaning that depends on the *intrinsic* parts of certain entities, e.g., people and cars have fronts and backs (see e.g. Miller and Johnson-Laird, 1976, Sec 6.1). It calls for a similar procedure based on the entities themselves rather than a line of sight.

In the deduction based on the premises:

The circle is in front of the triangle.

The cross is behind the triangle.

The first premise leads to the construction of a minimal spatial array that satisfies the truth conditions of the premise (assuming the appropriate viewpoint):

A  
O

The information in each subsequent premise can be added to the model, inserting tokens in the appropriate place in the array to satisfy the meaning of the premises. Thus, the second premise yields:

+  
A  
O

The conclusion:

The circle is in front of the cross.

contains only referents that are in the model, and so a procedure is called to verify the assertion in the model. It evaluates the assertion as true in the current model, and so another procedure is called to try to falsify it by finding an alternative model of the premises. This procedure fails in the current case, and so the conclusion is valid.

The premises:

The circle is in front of the **triangle**.

The cross is behind the circle.

yield the initial model:

+  
A  
O

that supports the conclusion:

The triangle is in front of the cross.

But, in this case, the falsification procedure succeeds in constructing an alternative model of the premises in which the conclusion is false:

Δ  
+  
O

The two models do not **support** any relation in common between the triangle and the cross, and so no valid deduction can be **made** about them.

Experimental results corroborated Clark's linguistic theory. But they are also consistent with the construction of a spatial array whose top represents the highest or positive end of a scale. Subjects prefer to work from the top down, and to construct an array from an "end-anchored" premise, i.e. one in which the first noun phrase refers to an item at one end of the array. Although imagery *per se* does not seem to play a large role in transitive deductions (Richardson, 1987), the evidence implies that subjects do construct a linear array of items (e.g., Barclay, 1973). Similarly, when the relevant entities are far apart in the array, subjects are faster to make decisions about the relation between them (Potts, 1978; Newstead, Pollard, and Griggs, 1986). They also make more inferences from sequences of conditionals that are transitive than from those that are not **transitive** — a phenomenon that again suggests that

they try to construct integrated representations (Byrne, 1989b). Nevertheless, the rule and model theories make much the same predictions for three-term series problems.

In response to the **impasse**, some theorists have proposed that reasoners use both rules and **models**, either at **different** points during the process (e.g. Johnson-Laird, 1972; **Sternberg** and Weil, 1980; **Sternberg**, 1985), or as alternative strategies (**Egan** and Grimes-Farrow, 1982; **Ohlsson**, 1984). A more radical way around the impasse is to examine problems for which the theories do make different predictions. A promising form of relational reasoning, which is fundamental to understanding the **world**—for planning routes, locating entities, and envisaging **layouts**—**depends** on two-dimensional spatial relations.

## TWO-DIMENSIONAL SPATIAL DEDUCTIONS

When people understand spatial descriptions, they imagine symmetrical arrays in which **adjacent** objects have roughly equal distances between them (Ehrlich and Johnson-Laird, 1982). They can represent these descriptions in two ways: one is close to the linguistic structure of the sentences, and the other is close to the structure of the situation that is described (Mani and Johnson-Laird, 1982). How then do they make deductions about spatial relations?

Consider the following problem:

I. A is on the right of B

**C is on the left of B**

D is in front of C

E is in **front** of B

What is the relation between D and E?

**Hagert** (1983, 1984) has proposed a rule theory that includes the rules shown in Table 5.1. In combination with rules for propositional connectives, they allow an answer to the question to be derived:

**D is on the left of E.**

and Table 5.2 presents the derivation.

Here is a second problem:

II. B is on the right of A

**C is on the left of B**

D is in front of C

E is in front of **B**

What is the relation between D and E?

**Table 5.1**

Some inference rules for one-dimensional and two-dimensional deductions (from Hager, 1983)

- 
- a. Left (x, y) & Front (z, x) → Left (front (z, x), y)  
where the right-hand side signifies "z is in front of x, which is on the left of y".
  - b. Left (x, y) & Front (z, y) → Left (x, front (z, y))  
where the right-hand side signifies "x is on the left of z, which is in front of y".
  - c. Left (x, y) & Left (y, z) → Left (x, left (y, z))  
where the right-hand side signifies "x is on the left of y, which is on the left of z".
  - d. **Left (x, y) ↔ Right (y, x)**
  - e. Left (front (x, y), z) → Left (x, z) & Left (y, z) & Front (x, y)
  - f. Left (x, front (y, z)) → Left (x, y) & Left (x, z) & **Front (y, z)**
  - g. Left (x, left (y, z)) → Left (x, y) & Left (x, z) & Left (y, z)
  - h. Left (x, y) → ¬ Right (x, y)
  - i. Right (x, y) → ¬ Left (x, y)
- 

**Table 5.2**

The derivation of a spatial deduction using Hager's (1983) rules

*The premises:*

- 1. A is on the right of B
- 2. C is on the left of B
- 3. D is in front of C
- 4. E is in front of B

**Hence, D is on the left of E.**

*The derivation:*

- 5. C is on the left of B and D is in front of C (conjunction of 2 & 3)
  - 6. D is in front of C, which is on the left of B (rule a applied to 5)
  - 7. D is on the left of B, and C is on the left of B, and D is in front of C (rule e applied to 6)
  - 8. D is on the left of B (conjunction elimination applied to 7)
  - 9. D is on the left of B & E is in front of B (conjunction of 4 & 8)
  - 10. D is on the left of E which is in front of B (rule b applied to 9)
  - 11. D is on the left of E, and D is on the left of B, and E is in front of B (rule f applied to 10)
  - 12. D is on the left of E (conjunction elimination, 11)
-

It can be solved using exactly the same derivation. The first premise is irrelevant in both **problems**, and the remaining three premises are identical. Hence, if people are using rules, there should be little **difference** in difficulty between problem I and problem II.

In contrast, the model theory predicts a difference. The premises of problem I support the model:

C	B	A
D	E	

D is on the left of E in this model, and it is a valid conclusion because the procedure that revises models cannot construct an alternative model to refute it. The premises yield a one-model problem with a valid conclusion.

Problem II supports at least two distinct models:

C	A	B	A	C	B
D		E		D	E

but both of them support the same conclusion:

**D is on the left of E.**

and no alternative model falsifies it. The premises yield a multiple-model problem with a valid conclusion. For one-dimensional problems, the validity of a deduction is confounded with the number of models that it requires: one model problems support a valid deduction, multiple-model problems do not. Two-dimensional problems, however, allow us to disentangle the two **variables**. And the model theory predicts that the second problem should be **harder than the first, because it** should be harder to make deductions based on more than one model,

A third sort of problem is exemplified by the premises:

III. B is on the right of A

**C is on the left of B**

D is in front of C

**E is in front of A**

What is the relation between D and E?

In this case, there is no valid answer; and both the rule and the model theories predict that this sort of problem should be hardest of all. According to the rule theory, it is difficult because all potential derivations have to be tried before one can respond that the problem has

no valid answer. According to the model theory, the problem is **difficult** because it requires at least two models to be constructed in order to appreciate that there is no valid conclusion relating D and E:

C	A	B	A	C	B
D	E		E	D	

It is a multiple-model problem with no valid conclusion about D and E, and so it should be harder than the one-model problem. It should also be harder than the multiple-model problem with a valid conclusion, where the valid conclusion emerges even if only one model is **constructed**—the conclusion is true in every **model** of the premises.

We tested the predictions of the two theories in an experiment in which 15 subjects made one-dimensional and two-dimensional deductions about the layouts of everyday **objects**, such as cups and plates (Byrne and **Johnson-Laird**, 1989b). The percentages of correct conclusions were as follows:

One-dimensional problems:	
Valid conclusion (one model)	69%
No valid conclusion (multiple model)	19%
Two-dimensional problems:	
Valid conclusion (one model: e.g. problem I)	61%
Valid conclusion (multiple model: e.g. problem II)	50%
No valid conclusion (multiple model: e.g. problem III)	18%

We expected that the one-dimensional problems would be easier than the two-dimensional ones, but there was no reliable difference between them. As both theories predicted, the valid problems were easier than the invalid ones. The crucial finding, however, occurred with the two-dimensional problems. The one-model problems with a valid conclusion were reliably easier than the multiple-model problems with a valid conclusion. This difference supports the model theory, and runs counter to the rule theory.

A stronger test between the two theories would pit them directly against one another, in cases where the rule theory predicts a difference in one direction, and the model theory predicts a difference in the

**opposite** direction. Our next experiment made **such a comparison**, using a **fourth** sort of problem:

- IV. A is on the right of B  
**C is on the left of B**  
 D is in front of C  
 E is in front of A  
 What is the relation between D and E?

This problem does not contain any premise that directly asserts the relation **between** the pair of **items**, A and C, to which E and D are respectively related. It is therefore necessary to deduce the relation between A and C from the first two premises:

1. **A is on the right of B**
2. **C is on the left of B**

These premises **yield**, according to the formal rules (see Table 5.1), the following sequence of inferences:

3. B is on the right of C  
 (modus ponens from 2 and instantiation of rule d)
4. A is on the right of B and B is on the right of C  
 (**conjunction** of 1 and 3)
5. A is on the right of C  
 (modus ponens from 4 and instantiation of **transitivity**)

This conclusion and the remaining two premises now permit an **analogous** derivation to the one for problem II, where B and C were the relevant **pair** of items and directly related in the second premise.

Problem IV requires a longer derivation than problem II, and so it should be harder according to the rule theory. Problem IV, however, supports just one model:

C	B	A
D		E

whereas problem II requires more than one model. Hence, the two theories make exactly opposite predictions.

We tested the predictions in a second experiment in which 18 subjects carried out three sorts of two-dimensional problem: one model problems **with** a valid conclusion (such as IV), multiple model problems with a valid conclusion (such as II), and multiple model problems with **no** valid

conclusion (such as III). The problems concerned everyday objects, and the layouts were in one of four orientations:

1			2			3		4	
A	B	C	E		D	D	C	A	E
D		E	A	B	C		B	B	
						E	A	C	D

The percentages of correct conclusions were as follows:

Valid conclusion (one model: IV)	70%
Valid conclusion (multiple model: II)	46%
No valid conclusion (multiple model: III)	15%

These results support the model theory, but not the rule theory. Despite the fact that the derivation for IV is much longer than the derivation for II, problem IV was reliably easier. Problems with no valid conclusion were, as both theories predict, hardest of all.

Both experiments show that it is easier to make a valid deduction when a description corresponds to just a single layout as opposed to two or more layouts. The phenomenon is explained if people reason by imagining the state of affairs described in the premises, drawing a conclusion from such a mental model, and searching for alternative models that might refute the conclusion.

Could our instruction to imagine the layouts have led the subjects to adopt a strategy otherwise alien to them? We used the instruction, which was casually mentioned during the subjects' introduction to the task, in order to avoid any problems in the interpretation of "on the right of", "on the left of, and the other spatial terms. Only the deictic sense yields the deductions of the sort used in the experiment, and a simple way to ensure that the subjects made this interpretation rather than the one based on intrinsic parts was to tell them that the layouts were being described from a particular point of view. It is unlikely that this instruction could be powerful enough to cause subjects to adopt a wholly unnatural reasoning strategy. Indeed, several authors have lamented the difficulty of inducing reasoning strategies by explicit instructions (e.g. Dickstein, 1978), while others have denied a significant role for imagery in relational reasoning (Newstead, Manktelow, and Evans, 1982; Richardson, 1987). The critical feature of the model theory is the structure of the representations used in reasoning —they should be the

same as the structure of the **world**, not the structure of **sentences**—rather than that they should be experienced as images.

### ALTERNATIVE RULE THEORIES FOR SPATIAL DEDUCTIONS

Our experiments are damaging for rule theories of the sort shown in Table 5.1. Could a theory with **different** rules account for our results? One alternative is worth describing because it illuminates the peculiar difficulties that confront rule theories. It postulates such principles as:

If x is on the left of y, and w is in front of x,  
and z is in front of y, then w is on the left of z.

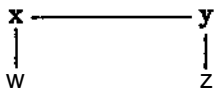
The first premise in problems I and II is irrelevant, but the application of this rule to the remaining premises yields the conclusion directly:

D is on the left of E.

Because a large number of such rules would be necessary to deal with the full set of spatial relations, we can invoke general postulates of the same structure:

If x is related to y on one dimension,  
and w is related to x on an orthogonal **dimension**  
**and** z has the same orthogonal relation to y,  
then w is related to z in the same **way** as x is related to y.

This general postulate has the advantage that it captures the content of many specific postulates, because it can be applied to any rotation or reflection of the configuration:



Yet, a theory based on this rule still makes the wrong predictions. It matches the premises of the difficult problem directly, where C is directly related to **B**:

- II. B is on the right of A
- C is on the left of B**
- D is in front of C**
- E is in front of B

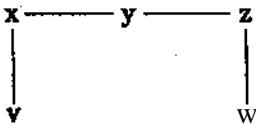
and so the second premise corresponds to the clause in the postulate:

If  $x$  is related to  $y$  on one **dimension**...

There is no such correspondence in the case of the easy problem, IV, where it is necessary, as we showed, to deduce the relevant relation (between A and C). Hence, the postulate predicts the opposite of the observed difference. One might be tempted to invoke a rule that does match problem IV directly, **e.g.:**

If  $x$  is related to  $y$  on one dimension,  
 and  $y$  is related to  $z$  on the same dimension,  
 and  $v$  is related to  $x$  on an orthogonal dimension,  
 and  $w$  is related to  $z$  on the orthogonal dimension,  
 then  $v$  is related to  $w$  in the same way as  $x$  is related to  $z$

which can be used for any orientation of the configuration:



Unfortunately, the rule does not match the premises of the other easy problem, I, and so now this problem should be difficult. In short, the three sorts of valid problem that we have studied put any rule theory on the horns of a dilemma. If the theory explains the ease of problem I it cannot explain the ease of problem IV, and vice versa. And, if the inferential system has all of the rules that we have considered, then no problem should be harder than any other.

## CONCLUSIONS

The model theory overcomes the difficulties for rule theories because it separates the representation of the premises from subsequent deductive processing. There are **many** possible models corresponding to the possible layouts of objects, but any deduction can be made by the same **process**—the search for alternative models that refute the relation established in a model. For rule theories, however, the process of deduction depends on the particular rules and postulates that are used in deriving the conclusion, and so different deductions depend **on**

different rules. The need for a **derivation's length**, in **addition**, to predict the difficulty of a deduction places an impossible load on the theory.

The model theory readily accommodates such deductions as:

The door is taller than it is broad.

The table is wider than the door is **tall**.

Therefore the table is wider than the door is broad.

Reasoners can construct a model of the situation from their knowledge of the meaning of the relational **terms**, and they can compare the relative sizes of the objects in any model that satisfies the premises. To derive the conclusion using formal rules, however, calls for a complicated procedure. It is necessary to use postulates that interrelate such predicates as "**broad**", "**wider**", and "**taller**", and that enable the relative sizes of **different** dimensions to be compared.

Critics sometimes argue that the use of spatial models smuggles in inference rules by the back door. But, the procedures in our computer program that operate to construct and to inspect arrays are certainly not formal inference rules, which, by definition, lead directly from verbal premises to verbal conclusions. Likewise, the representations of the meanings of relations are not meaning postulates that **specify** logical properties. The essence of the semantic entry for "on the right of", for example, merely ensures that one object is on the right of another in a model (see Chapter 9 for the details of how such meanings can be represented). The meaning does not in itself specify that the relation is transitive, but this logical property emerges as soon as the meaning is put to use in building models.

The same point can be made by considering the definition of "**greater than**" in the theory of recursive functions. One **number**,  $x$ , is greater than another,  $y$ , if and only if  $x$  is the successor of  $y$  or the successor of some other number,  $z$ , that is greater **than**  $y$ . This recursive definition depends only on the concept of the successor of a **number**—a mathematical function that, given a natural number, such as 5, delivers its successor: 6. When you learn to count, you master this function. It is in no sense transitive, i.e. if  $x$  is the successor of  $y$ , and  $y$  is the successor of  $z$ , **then**  $x$  is *not* the successor of  $z$ . Nevertheless, the transitivity of "greater than" is an emergent property of the recursive definition. The general message of recursive function **theory** is that the richness of computable mathematical functions and relations is likewise emergent from different assemblies of a handful of simple building blocks (Rogers, 1967).

Other critics have suggested that the use of arrays trades unfairly on a visual metaphor. Our computer program is unable to exploit the

metaphor and yet it is certainly capable of modelling the deductive process. The construction of a model depends on understanding the meaning of a relational term. The logical properties of the relation then emerge from its use in constructing models. Indeed, there is an important asymmetry between the meaning of a relation and its logical properties. Logical properties can emerge, as they do in our program, from meaning. But, meaning cannot emerge from logical properties. Many distinct relations have identical logical properties, e.g. both "on the right of and "on the left of are transitive, asymmetric, and **irreflexive**. No matter how many other logical properties a rule theory adduces, it will never be able to distinguish between the meanings of these two relations.

Our aim in this chapter was to reach a conclusion about how people make relational deductions. We have argued that they do so by imagining the state of affairs described by the premises. Rule theorists may indeed concede that spatial deductions are based on mental models rather than formal rules. We will show in subsequent chapters that relational reasoning is fundamental to many abstract deductions, and that the model account extends to them too.