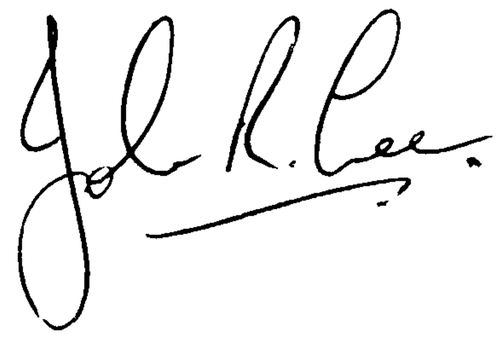
Metalogic
and the
Psychology of Reasoning

John Richard Lee

Ph.D.
University of Edinburgh
1987

Declaration

I declare that this thesis is my own work,
and has been composed entirely by my own hand.



November 1987.

Abstract

The central topic of the thesis is the relationship between logic and the cognitive psychology of reasoning. This topic is treated in large part through a detailed examination of the recent work of P. N. Johnson-Laird, who has elaborated a widely-read and influential theory in the field. The thesis is divided into two parts, of which the first is a more general and philosophical coverage of some of the most central issues to be faced in relating psychology to logic, while the second draws upon this as introductory material for a critique of Johnson-Laird's 'Mental Model' theory, particularly as it applies to syllogistic reasoning.

An approach similar to Johnson-Laird's is taken to cognitive psychology, which centrally involves the notion of computation. On this view, a cognitive model presupposes an algorithm which can be seen as specifying the behaviour of a system in ideal conditions. Such behaviour is closely related to the notion of 'competence' in reasoning, and this in turn is often described in terms of logic. Insofar as a logic is taken to specify the competence of reasoners in some domain, it forms a set of conditions on the 'input-output' behaviour of the system, to be accounted for by the algorithm. Cognitive models, however, must also be subjected to empirical test, and indeed are commonly built in a highly empirical manner. A strain can therefore develop between the empirical and the logical pressures on a theory of reasoning.

Cognitive theories thus become entangled in a web of recently much-discussed issues concerning the rationality of human reasoners and the justification of a logic as a normative system. There has been an increased interest in the view that logic is subject to revision and development, in which there is a recognised place for the influence of psychological investigation. It is held, in this thesis, that logic and psychology are revealed by these considerations to be interdetermining in interesting ways, under the general a priori requirement that people are in an important and particular sense rational.

Johnson-Laird's theory is a paradigm case of the sort of cognitive theory dealt with here. It is especially significant in view of the strong claims he makes about its relation to logic, and the role the latter plays in its justification and in its interpretation. The theory is

claimed to be revealing about fundamental issues in semantics, and the nature of rationality.

These claims are examined in detail, and several crucial ones refuted. Johnson-Laird's models are found to be wanting in the level of empirical support provided, and in their ability to found the considerable structure of explanation they are required to bear. They fail, most importantly, to be distinguishable from certain other kinds of models, at a level of theory where the putative differences are critical.

The conclusion to be drawn is that the difficulties in this field are not yet properly appreciated. Psychological explantion requires a complexity which is hard to reconcile with the clarity and simplicity required for logical insights.

Contents

Preface

The nature of reasoning is one of the most profound questions to be faced by cognitive science. We are faced with a task, the description of which is evidently drawn from the field of logic, but yet the performance of which must be furnished with a psychological account. In the very notion of 'inference', we find components, drawn from these two fields, which at times sit uneasily together. What has to be examined is the relationship between them.

This is not a new problem. All approaches to the subject of psychology have had at some stage to grapple with it. Attempts have ranged in nature from the austerities of behaviourism to the epistemological convolutions of Piagetian theory. The temptation has often been, on the one hand, to emphasise the similarities between the reasoning task and any other, while neglecting all but entirely the relevance of its logical description, or on the other hand to fail in recognising the complexity and richness of the behaviour engendered by reasoning tasks, through overplaying the formal element. It is through the reconciliation of these two aspects that cognitive science will contribute to progress in this area, if indeed it will at all. We should therefore seek to evaluate the opportunities it offers.

The idea of a cognitive model is what we must look at. In constructing such a model, the cognitive scientist is attempting to create a formal framework within which to produce an explanation of observed behavioural phenomena. If anything can unite the two apparently opposed approaches to the study of reasoning, then surely this is it.

This is not, of course, something that is going to be accomplished all at once, or in a single work. However, there have already been some salient attempts to get to grips with the issues involved here. In the course of this thesis, we shall be devoting a major part of our attention to just one of these, *viz.* that presented by P N Johnson-Laird, in *Mental Models* (1983) and elsewhere, particularly as it applies to syllogistic reasoning. This particular theory is chosen because it exhibits several important characteristics: it is expounded in considerable detail; it is explicitly directed at deductive reasoning; it makes very strong

and highly interesting claims, about just the topics we have mentioned; it is a widely-read and influential theory at present in the field. A detailed examination of Johnson-Laird's theory will lead us at the same time to examine the fundamental problems which any such theory is going to have to face, and will therefore be of considerable value.

* * *

This thesis falls into two parts. In the first part, there is a relatively philosophical discussion of some of the more general issues involved in the field. The intention here is to provide an orientation and some introductory material for the more detailed aspects of the second part. There is some consideration of the ways in which reasoning tasks should be identified, for the purposes of theorising about them; the suggestion is that such identification has not always been done very appropriately in the past. A particular attitude towards the idea of a cognitive model is outlined, and questions then addressed about how this relates to matters such as competence, normativity and rationality. Throughout this part, reference is made to the work of Johnson-Laird, because this importantly connects it with the second part. The second part itself is a detailed examination of the theory offered by Johnson-Laird in 1983, attempting both to relate it back to the issues of the first part, and also to consider some of the problems it raises specifically, for instance in respect of the extent to which it shows a proper account of mental processing essentially to involve semantic notions, and the extent to which its claims are really susceptible to empirical test. We confine our detailed attention almost wholly to *deductive* reasoning, mainly because it has received the most detailed treatment at the hands of specifically cognitive theorisers. Inductive and probabilistic reasoning has been extensively studied by social psychologists interested in rationality, and consequently it does receive significant mention in Part I.

The hope is that we can achieve three things: firstly, the philosophical examination of some very general and highly important issues about the relationship between logic and cognitive psychology; secondly, by treatment of a specific example, a demonstration of the ways in which these issues appear in the psychotheoretical practices, as it were, of actual cognitive psychologists; and thirdly, a detailed dissection of a current theory of considerable interest and importance in the field, which would be worthwhile in its own right. Although an attitude to the topic is adopted and argued for, it will be clear that this is a thesis which takes more the form of a critique of certain current ideas than that of a proposal of some particular theory of its own. Its contribution to the field is accordingly to be sought chiefly in the improved understanding of the problems in the field, which it claims thereby to bring.

An expositional strategy employed frequently throughout is the close and critical examination of a paper or similar text from the relevant literature. Such texts are carefully chosen to provide an interesting and significant treatment of the matters we are interested

in, so that they provide a natural starting point for consideration of wider issues. This strategy is perhaps mostly the result of personal preference, but it is supposed also to anchor the discussion within the frame of reference provided by previous work. I have not included a separate and comprehensive literature survey, mainly because it seemed redundant in view of the excellent ones available elsewhere, and I felt that the space could be put to more interesting use. The interested reader can do no better than to look at the surveys provided:- on deductive reasoning, by Evans (1982), and Johnson-Laird (1983); on inductive and probabilistic reasoning, by Nisbett and Ross (1980); on inductive logic, by the monumental work of Kyburg (1983); on the general issues of rationality, by Elster (1982) and various of the contributors to Geraets (1979).

•

Acknowledgements

Thanks are due to a number of influences which have improved my understanding of the issues addressed in this thesis. I gained a great deal from discussions with the participants of a workshop on mental models in inference, held at the School of Epistemics some years ago, including Robert Inder, Terry Myers, Han Reichgelt and my supervisor, Barry Richards. Some of the members of this group were involved also in a later workshop to which I am likewise grateful, especially for encouragement from Brendan McGonigle and Maggie Chalmers. I have tried to indicate all occasions where I draw on the ideas of any of these people, and I apologise to anyone who has been overlooked.

Above all, this thesis owes its existence to the patience and sacrifices of my wife, Rosemary, and our children, who have grown up along with it.

Part I

1

Introduction to Part I:
Rationality and Reasoning

The subject of the following discussion is the cognitive psychology of deductive reasoning. The intention is to clarify certain issues about how this area in psychology is affected by logic, and (what may surprise some) how it affects logic, and what the relevance of this is to various problems which arise about rationality. My claim is basically that one's theory of cognitive psychology, if it is done seriously, and one's views about logic, are strongly mutually determining. That is to say, the cognitive psychology of deductive reasoning is a point where logical issues have empirical consequences, and where empirical results bear most seriously on logical matters.

1.1 Rationality and Logic

A frequent starting point, among those who study reasoning, is that it is to be regarded as a process of drawing inferences. It is further supposed that these inferences can, at least in principle, be characterised by some normative system of logic (eg. the first-order predicate calculus). It is therefore concluded that certain inferences are *correct* - *viz.* the ones which are logically valid - and that ideal rationality consists in the drawing only of such inferences.

The view is infected with certain difficulties, which stem from the philosophy of logic and from what some might insist is a slightly different subject, 'metalogic'. In many cases, this can be harmless; often, when one uses in one field some ideas or results derived from another, one can afford to overlook many niceties that would trouble experts in that other field - and this case is no exception. Sometimes, though, one can tread too closely to the interests of the field from which one poaches. The issues which one addresses become hard to separate from the issues addressed there, and then one has to take account of the arguments in that field, and what they may entail for one's own pursuit. This is what can happen when psychologists interested in reasoning proclaim results or theories which are

relevant to fundamental issues in the study of logic itself.

This is not to suggest that such overlapping is necessarily bad, nor that it always (or even generally) goes unnoticed by its perpetrators. Sometimes psychologists insist that their results are relevant to the studies of philosophers and logicians, and that the latter cannot afford to ignore them; at other times, claims are simply made which, when brought to the attention of the philosopher or logician, cry out for close scrutiny. Whether deliberately thrust before us or not, these are the kinds of claims that motivate the present attempt to uncover what the presuppositions of the psychology of reasoning actually are, with respect to logic and rationality, and contribute to the determination of their consequences.

Rationality is a difficult thing to characterise. Many, indeed, have tried, and abandoned the effort. I don't propose here to get involved in trying to define it in any sort of detail, or with any sort of clarity, although we shall examine it in some detail later. However, I want to insist that it is an error to identify rationality with conformity to the norms of some canonical formal theory, eg. classical propositional logic, or mathematical probability theory. We shall presently see some particularly clear cases of this error in action; cases where the perpetrators are quite open about what they are doing (although, of course, not regarding it as misguided). The idea that it really is an error depends on a certain approach to logic, and other formal theories, and the matter of how they get justification for their normative pronouncements. My belief here is that such justification in fact depends on the assumption that the theory captures a system of rational inference. This appears consistent enough with what I have just condemned, but the point is that the justification flows, as it were, from rationality to logic, and not vice versa. Insofar as a theory's characterisation of certain inferences as 'correct' has any normative force, this derives from the presumed universality of the injuction: *Be rational!* The fact that an inference is not sanctioned by a particular logic only shows that inference to be even a possible symptom of irrationality if it is granted that the logic adequately captures rational inference.

This, however, is always something that may turn out not to be the case. Many, even some of those most guilty of the error just described, concede (nay, declare) that logic is open to revision and progress (*cf.* the discussion of Nisbett's work in chapter 5, below). But what could possibly lead to any such revision, save the realisation that some inferences previously sanctioned were not, after all, within the set of those warranted by considerations of the rationality of the system; or that some, formerly abrogated, ought now to be indulged. An inference's being outside a particular logic, therefore, may only show that it is, in a manner of speaking, ahead of its time; that its rationality is not yet recognised or formalised.

1.2 Rationality and Psychology

A tendency has arisen among certain psychologists in recent years, most particularly social psychologists interested in inductive reasoning and statistical judgement, to make claims concerning human rationality. '*Is man a rational animal after all?*', they ask. Experiments are designed which purport to be able to answer this question empirically, by finding out, for instance, whether people's probability judgements are commonly, or even often, remotely correct, according to the canons of probability theory. Similar questions are addressed concerning deductive reasoning: can people correctly infer the conclusions to syllogisms, say, or even fairly simple propositional arguments? If it's the case that people are typically wrong in these sorts of tasks - if people are systematically irrational - there are consequences, it is claimed, for metalogical theories, especially those which address the justification of inferences. Goodman's ideas about 'reflective equilibrium' are a case in point: if people are irrational, how can some judgement's being held by them to be correct go very far towards justifying it?

There are two points to be investigated about this line of argument. One is the extent to which experiments of the sorts which are generally adduced (or indeed of *any* sort) can in fact establish such a conclusion as that people are *irrational*. (Does the fact that people persistently err actually show this?) The other is the size of the impact that whatever conclusions can be thus established might have on various logical and philosophical issues. We have said that justification and rationality are very closely linked: the point of a logic is to capture the notion of rational inference. There are connections here with the claim often made, especially in the context of theories of radical translation, by such as Quine, Davidson, Dennett, etc., that intentional descriptions of organisms - ie, descriptions in terms of *beliefs*, *desires*, and so forth - depend centrally on the assumption that the subject *is* rational.

If we take it for granted that formal theories in general, and logics in particular, are revisable, the question naturally arises as to the circumstances in which this may come to be seen as necessary. The disposition is strong, to assume that one's logic actually does faithfully reflect the pattern of rational inference, and to repulse all attempts at refutation. People's intuitions about what is rational do not always coincide, but they seem to have a tendency to fall into discernible groupings, which then give rise to different systems of logic existing at the same time and, to some extent at least, in competition. This happens, for instance, with the classical and intuitionistic systems of propositional logic. These both seem quite stable, and both command a following, but they disagree on such fundamental points as whether the disjunctive syllogism (whereby it follows, from *A or B* and *not A*, that *B*) is a theorem.

One possibility is that two different systems may both be applicable in certain circumstances, but not at the same time. Thus, it has been suggested that in quantum physics, for instance, some non-classical logic may best handle the kinds of inferences that seem to have to be allowed. Still, a peculiar domain like that carries little compulsion to suggest that some non-classical logic may be appropriate in any perfectly ordinary, homely field of everyday inference - which is clearly what would be required for a serious contemplation of revising standard reasoning norms. Nonetheless, there is interest in the idea that different logics may apply in different cases, and it will become significant later.

* * *

What I want to suggest at the moment is that if there is a possibility that standard norms should be revised, the cognitive psychology of reasoning is a good place to look for an indication. It should be declared here that I am in this thesis assuming a particular view of cognitive psychology, laid out in some detail in chapter 3 below, which is very similar in spirit to the views of Johnson-Laird, as will emerge in Part II. This view, of which I am concerned to examine the consequences rather than construct a defence, entails that cognitive psychology depends on the construction of a cognitive model of processes - in this case, reasoning processes, the drawing of inferences - and such a model requires the postulation of rules (or at least rule-governed processes) constituting a central algorithm showing (inter alia) what conclusions will, in ideal cases, be given to which premises. This central algorithm constitutes an explication of the *reasoning competence* of the system it models. The common failure to follow these rules to the letter has to be explained in some relatively *ad hoc*, and (as it will appear) in a cetain sense *non-cognitive* fashion, as for example by the supposition that various unprincipled effects such as memory limitations disrupt the smooth workings of the system.

An often overlooked consequence, here, is that the construction of a cognitive model implicitly involves the inclusion of a normative system. Statements are effectively made about what people are capable of, and the view is rarely contested that, since 'ought' implies 'can', this influences what one may say about how they should think or act. Given a set of data about people's actual performances on a particular task, there is always a *decision* to be made about which of these are *competent* performances, and which are not. This is something that depends entirely on the theory being developed to explain the performances, and cannot be divined from them in isolation; it is something which, in a certain sense, has to be contributed by the theoriser in an *a priori* fashion. Or at least, partly *a priori*: the qualification appears advisable because, after all, it is plausible to suppose that the theory is constrained, for instance to account for the performances in a reasonably *economical* way, and that the fulfilment of these constraints is evaluated with respect to something empirical. The main theoretical crutch, eg. for the psychologist who holds that

he can demonstrate irrationality, has to be that he has a theory which explains the data, using the irrationality assumption, somehow in a better, perhaps in a simpler, way than one which avoids it. Supposing that people have rational competence, he argues, requires too many unprincipled complexities to 'save the phenomena'; his theory is in some sense closer or truer to the performances.

1.3 The Role of Competence

Another common tendency among psychologists, is to assume from the outset that ideally competent reasoners are ideally rational, and therefore that this competence is identifiable with classical propositional logic, regarded as definitive of rational inference. However, if psychology is an empirical science, one supposes it surely must be allowed as possible that people's competence is not, in fact, equivalent to this logic. This is something that psychologists might discover; it might turn out that no model implicating such a competence can be made to give a plausible account of reasoning processes. Such a result could be regarded as having one of two consequences: either people are not rational, or the logic does not after all successfully capture rational inference. On what grounds would one decide between these?

While there are those who think that such a discovery does indeed show humans to be irrational because their competence is not equivalent to a specified norm, there are equally those who seem convinced that human competence actually *must* be equivalent to some favoured logic (usually the one mentioned above), and thus hold some particular characterisation of competence to be a condition of the plausibility of any cognitive reasoning model. Such a position is always possible, because (although it may strain the credibility of other parts of the theory) one can always account for *any* given performance data using a given competence, if one builds in enough *ad hoc* provisos about memory limitations, etc., that prevent the competence from being 'realised'.

One can yet take the view, alternatively to either of these positions, that, while people evidently do not have the competence of a given logic, this shows not so much their irrationality, as the inadequacy of the logic in question. There is thus a curious triangular trade-off between one's attitudes towards rationality, logic, and competence, depending on the kind of intuitions one has about them. The intuition that they ought all to be in some sense equivalent, by itself drives a tendency to respond to empirical discrepancies by revising logic and competence theories. The intuition that the first two are already well-defined and fixed, leads similarly to the assertion that people can be shown to be irrational, and have no fully logical competence. But these intuitions, when combined together, can also result in the apparatus of excuses to account for performance errors.

In what follows, we shall examine this trade-off, and, being generally sympathetic to the first-mentioned intuition, we shall find support for our view about the interdependence of logic and cognitive psychology in various arguments concerning the justification of deductive inferences. These questions about rationality are clearly embedded in a considerable tangle of often only half-recognised presuppositions, encompassing issues ranging from the logical to the methodological. The attempt to sort these out seems interesting and useful in its own right, but is further motivated by the examination of Johnson-Laird's theory of reasoning, which occupies the second part of this thesis. Johnson-Laird's theory is one which depends particularly crucially on certain notions about competence and rationality - notions which, to his great credit, Johnson-Laird goes to some lengths to set out and defend. They remain, nonetheless, a potent source of difficulties, and on them depend several of the claims made for the far-reaching significance of his theory. We shall find it useful to discuss several of these aspects of his ideas during the course of this first part, and it will be found that these efforts will pay a dividend in the second.

2

Reasoning Tasks and Logical Form

An obviously important question, but one rarely addressed in the detail it deserves, concerns what the psychology of reasoning is actually about. What is a theory of reasoning a theory of? We ought to examine the methodological background against which such theorising occurs.

2.1 What is Reasoning?

A natural first step, is to take reasoning to be the drawing of inferences. But evidently not just any inferential performance counts as reasoning, since, as cognitivists are forever reminding us, virtually all of cognition can be seen as involving inferential processes. It requires countless inferences to see an object, comprehend a sentence, pot a snooker ball, etc. A theory of reasoning has to be distinguishable from theories - even cognitive theories - of everything else.

The field is commonly restricted to linguistic cases, ie. where the premises of the inference can be found in some piece of language taken in by the reasoner, who then expresses his conclusions linguistically. But this is too restrictive for, eg, Johnson-Laird whose archetypal reasoning-situation, presented as a fundamental challenge for cognitive science, involves a conclusion expressed in the non-linguistic behaviour of the reasoner. The example concerns an individual who, upon asking 'Where is the university?', and being told that some of a nearby group of people are from there, goes at once to ask them. On the basis of observing this behaviour, we postulate that the individual has inferred; as Johnson-Laird puts it:

> [The] behaviour depends on a chain of inferences that includes at its centre the following deduction:
> Some of those people are from the university.
> Any person from the university is likely to know where the university is.
> Therefore: Some of those people are likely to know where the university is.
> (Johnson-Laird 1983 23)

One can easily go further than this, and imagine a situation in which no linguistic behaviour occurs at all - for instance, the potting of a snooker ball as mentioned above. In such an example, we may be able to get a clearer view of what it is that we are interested in, when we talk of reasoning, so let's start by assembling some intuitions about what is going on here. Whereas many inferential processes are no doubt required at some (relatively low) level, just to point the cue in the right direction, draw it back, and then hit the ball, these are not generally the processes we're after. These we might want loosely to characterise as perhaps 'motor' skills, in opposition to the (relatively high-level) 'cognitive' operations we want to investigate. It is when the snooker player is considering his shot, integrating it into a strategy for the break, deciding which ball to pot, and where the cue-ball ought to go afterwards, that he is most likely to be reasoning in the sense we require.

But there is a curious grey area in between these extremes. To what extent we want to say that he reasons actually in executing the shot may depend on his level of skill - on the extent to which his actions have become 'automatic', or 'unthinking'. Does he *reason* that he needs backspin, or that he should hit the cue-ball just at this point, and just this hard? Does the cricketer *reason* that in order to clip the ball through wideish midwicket, he needs to angle the bat just so?

We might propose that the snooker player reasons in a particular case as follows:

> Reds still in the pack cannot safely be potted.
> *This* red is still in the pack.
> Therefore: *This* red cannot safely be potted.

Supposing that we wanted to write a snooker-playing computer program, for some kind of simulation purpose, we can imagine that this is the sort of rule we might apply in considering each ball, and perhaps assigning it a weight in favour of selection as the one to play. It seems less easy to think of the human player (at least, one of any ability) as functioning in quite this way, because we suppose that he in some sense 'intuits' that balls in the pack are unsafe, and in fact probably never really *considers* them at all. At this point, the cognitivist tells us that, although neither we nor he may be aware of it, the fact remains that such processing must be going on. But accepting this does not solve our problem, for we still may ask whether it is reasoning that he is doing, or something at a level lower than we want to look at. No doubt much of the most basic motor-control processing could in principle be formulated or modelled as sets of syllogisms: this shows neither that it is reasoning, nor that any explicitly syllogistic rules, representations or inference-algorithms are actually there involved. Somehow, we feel that in a case where an agent 'just knows' something, he has not *reasoned* his way to it on this occasion. He may have done so in the past; perhaps he no longer needs to: the intuition appealed to here is that if one relies, say, simply on memory of results previously worked out, one is not reasoning one's way

to them any more.

The difficulty we are having here is not, of course, confined to nonlinguistic cases. Practically all uses of language involve something that might be regarded as being in some sense reasoning. A theory of reasoning, however, surely ought not simply to subsume the whole of psycholinguistics. There must be a very close connection, and clearly one cannot reason about a discourse, say, without having understood it (or misunderstood it) - but can one not understand it without at the same time, or thereafter, reasoning on it? Should a theory of reasoning include those undoubtedly inferential processes which must be involved in such matters as resolving anaphoric references? Traditionally, the answer is often yes, but I think we should question this. It's easy to suppose that anything appearing clearly to be an inference is an instance of some more or less cognitively universal phenomenon, but it can become a difficult position to defend. Reasoning, it might be safer to say, is something that really takes off only after the premises have been interpreted. At least we can then identify the premises and conclusion relatively securely. This may seem to be a matter mainly of terminology, but we shall return to it in more detail after the following digression.

2.2 Basic Inferences: Johnson-Laird's Early Theory

To expose the kind of presuppositions that do tend to occur in psychologists' writing, let us here take time to consider in some detail part of an earlier theory offered by Johnson-Laird (1975). The point is not essentially to belabour Johnson-Laird for what he said then; it's worth noting that most of Johnson-Laird's views described below are retracted in his later work, and this is something we shall return to. The point is rather to find something more general to say about the psychological treatment of reasoning, and especially to indicate the treacherous ease with which it is possible to assume that some single simple mechanism can be described, and even modelled by a simple computer program, which covers all, or most interesting, cases. However, this will be a particularly worthwhile exercise, because it does relate to Johnson-Laird's later theories which are considered in the second half of this thesis.

In his paper, Johnson-Laird says, he is interested in discovering the 'set of psychologically basic inferences', by which he means those which cause no difficulty to 'logically naive subjects'. These terms are typical of those that psychologists use in describing the objects of their investigations, so we will be repaid if we pause, first, to consider what they cover.

For one thing, 'psychologically basic inferences' apparently have to be *couched* (for want of a better word) in a particular manner. A system, Johnson-Laird says, which eschewed most of the connectives in favour of negation and disjunction, would be

psychologically implausible. Intuitively, something like *modus ponens*, say, ought not to be recast in such terms. Further, 'an inference schema can hardly be considered basic if most people are incapable of carrying it out, or can only do so in a matter of minutes, subsequently giving a detailed resumé of a whole chain of deduction they have carried out to make the inference' (17). These inferences must not result from the combining of several other inferences.

Now, what becomes evident here is that we need to know whether we're talking about inferences that actually *do* get made, in ordinary language, or those that *could*. Johnson-Laird appears ambivalent about this. For instance, he gives as example 'inference schemata', *A therefore A*, and *A or A, therefore A*. Clearly, these sorts of inferences could be made, but it seems odd to suggest that they very often are. The fact that their *logical structure* is basic, in some sense, can be neither here nor there, psychologically speaking. Johnson-Laird says he's talking about 'inferences expressed in natural languages', but it is rarely indeed that one will hear natural sentences expressing propositions of the above forms (or rather, perhaps, sequences of them representing arguments of these forms), and even when one does, it is unconvincing to suggest that they represent *inferences* ('OK, if it's not a fish, it's not a fish!'). It turns out that Johnson-Laird actually regards these and their like as 'auxiliary inferences', not drawn, as it were, in their own right, but only to support subconclusions of more plausible inferences. Even as such, though, it is unclear why they would in general be needed. Johnson-Laird goes on to give other examples, but he finds a problem with some of these, in that we apparently need to 'curb their productivity' somehow, as otherwise we get things like

(1) Boys eat apples and Mary threw a stone at the frog.

which is 'barely acceptable'. Here is a case of a basic inference (derived from 'Boys eat apples' and 'Mary threw a stone at the frog') which *could* be made, but in fact never would be. 'Why not?' asks Johnson-Laird. His answer, it seems to me, results from confusing the question of the nature of the mechanism which produces sentences having a certain sort of logical structure, with the question of how best that logical aspect of them is to be analysed and related to the same aspect of other sentences. No doubt (1) doesn't get *produced*, but this isn't because there are restraints on a logical inference schema; rather, it's for extra-logical, but perhaps linguistically important, reasons. The suggested inference to (1) is a *perfectly good inference*; an inference which in some sense *would* be made, if the conjuncts ever were presented together. The fact that they never are has to do with the information-conveying nature and purposes of language, and the structure of the actual world which is conversed about; its explanation should be related to that of the fact that the thing is indeed barely acceptable to begin with.

In many ways, Johnson-Laird seems to have drawn a somewhat arbitrary line around a class of phenomena, some of which appear curiously out of place, and called them 'inferences'. In particular, there are examples like (1): how far *should* we regard some of the processes involved in language production as being fair game for a theory of inference, rather than as more appropriately to be addressed by a separate (for the immediately foreseeable future) theory of psycholinguistics? The connective 'and' is curious, in that it can often be regarded as implicit even where it's not present. For example, one might produce several declarative sentences in succession, or one might conjoin them with 'and's to produce only one. To a large extent, this is surely an arbitrary choice, but what's interesting is that similar constraints apply to what follows what in either case. One wouldn't say: 'Boys eat apples. Mary threw a stone at the frog.', any more than one would utter (1), and, as suggested above, for much the same reasons. But one might say: 'The sun is shining. It's a lovely day.', or: 'The sun is shining and it's a lovely day.', without much preference for either style. It's hard to see that moving from one to the other constitutes an *inference*. Can we really conceive of simultaneously apprehending the conjuncts of something like (1), without in an important sense apprehending the conjunction itself? Have we similarly to suppose that when we know both that 'it's raining' and that 'the sun is shining', we have firstly to infer the conjoint truth of these propositions, before we can decide which connective most appropriately to interpose between them (eg. 'but')? What is needed here is not so much anything to do with inference as an explanation of why certain sequences of declarative utterances are preferred in ordinary discourse. No doubt such a theory will postulate the existence of inferential processes, but these are again at a level distinct from one where a theory of reasoning as such should be expected to apply.

Surely, an account of these phenomena will have to be in large part *content driven*, for the explanandum is how assertions are made, which clearly depends on what is asserted, and why. If this is right, though, Johnson-Laird's 'auxiliary inference' machine must be inadequate at best, but perniciously misleading at worst. Even if the device produces all the correct behaviour, it's hard to see it as doing so in other than a rather *ad hoc* sort of way. Johnson-Laird has now redeemed himself by admitting precisely this (Johnson-Laird 1983 35), but it remains a type of argument not uncommonly seen in cognitive theorising.

In many respects, Johnson-Laird's arguments against 'mental logic' (which we review later in chapter 12) are relevant here, but although turning to content-sensitivity is desirable, it is not enough to eradicate the difficulties we have been looking at. The proposed field of coverage of a theory of reasoning, even one involving a fairly detailed mechanism, is all too often far too ambitious. It's noticeable that in his later theory Johnson-Laird abandons the short-term goal of giving the full details about propositional inference. He provides an algorithm, but points out that it works only for those relatively infrequent cases

where language is used truth-functionally, and moreover that it is empirically untestable (Johnson-Laird 1983, ch. 3, *passim*). In fact, the field he offers to cover is if anything even greater then it was previously, but he can only give brief, intuitive, anecdotal sketches of how he might achieve this.

* * *

Let us continue to pursue Johnson-Laird's earlier theory, and consider how he then thought to submit it to empirical test. The characterisation of 'basic' inferences is important here, since it is upon these that the account of reasoning is based. They are to be distinguished from more complex ones of which 'a detailed resumé can be given', but are otherwise not easily recognisable. One might seek to describe this as a distinction between inferences drawn, as it were, *subconsciously*, and those drawn more reflectively. Johnson-Laird says that 'the methodology of making such tests of the model [as whether it performs in the *same way* as human reasoners] is not obvious, although human retrospections are likely to provide some useful evidence'. Perhaps the idea is that they are aware of following a chain of reasoning composed of links that they cannot further describe - the basic inferences.

But in any event it's unclear how we should take this. In cases where human reasoners are provided with specific tasks, like 'draw an inference from these premises', we might expect to get some sensible retrospective answer to the question how the task was done. To take one of Johnson-Laird's examples, one might give the premises

(2) A and C
If A then B

expecting the answer 'B', and hope to get some such retrospective protocol as: 'Well, if A then B, and we know A because it says A and C, so B follows'. But this isn't quite how Johnson-Laird describes the matter: he talks of premises *of this form*. Here again, though, one is tempted to suspect that the process of inference is affected as much by the content as by the form, if content there is. Consider, for instance, the following snatch of dialogue:

(3) - Fred's coming and George is coming;
and if Fred's coming then Mary'll be coming too.
- Well I hope she's not a vegetarian then!

In this case, it looks as though the second speaker must in some sense have made an inference from premises of exactly the form of (2) (as well as at least one other, possibly syllogistic, inference that we might guess at). However, suppose you interrupt the people, who perhaps are planning a small dinner party, and ask the second speaker *how* the conclusion was arrived at. What will you get? I suggest: nothing of interest. At best the

speaker will produce a protocol not altogether unlike the one for (2) - but if he's more honest, he'll probably say, 'I don't know; it was just obvious'. In the first case, it's just a concocted story - a hypothesis or theory, certainly not seriously remembered thoughts - while the second reply gets us nowhere. It's unlikely, as well, that any average English speaker ever experiences the slightest difficulty in (3)-type situations, so does this give us reason to suppose that (2) represents a psychologically basic inference schema? Johnson-Laird's account requires that it is not, but the only criteria offered for basicness are lack of difficulty to the average speaker, and lack of structure. It might well, of course, be held that (2) - and hence (3) - has quite significant *logical* structure, but what has to be shown is that this results in any loss of *psychological* immediacy.

I think, in fact, that Johnson-Laird's criteria for basicness can be interpreted so as to be quite plausible, but on that interpretation (3) is basic, although (2) might not be. What we learn from this is that basicness is in some sense context-relative and cannot be determined by logical form. There is a similarity between (2) and (3), and in general between many formal representations of arguments and various sequences in discourse. However, in most cases, there is little *prima facie* evidence that this resemblance is anything other than spurious and misleading.

It's clear that Johnson-Laird's theory here involves at least a gross oversimplification of the range of inferential performances, and that there's no real way of showing that it applies to many of these cases where inferences are relatively 'tacit'. In some yet mysterious way, people have the ability linguistically to construct descriptions, etc, of the world: doing this in itself no doubt involves manifold inferences, yet surely on a level different from one where we might want to deploy rules and schemata in its explanation. This intuition is one which Johnson-Laird seemed to share, even when he wrote the paper under discussion. He says, for instance, that inferences which involve *combining* information, require his theory of mental models - and it's unclear why he thinks (3) is not such a case. A major feature of his later theory is in fact that it extends the idea of mental models to cover propositional inferences, rather than just those involving quantifiers as previously. In any event, an account of (3) in terms of rules and schemata would still seem to leave many other, still more tacit cases unaccounted for. It is more plausible in some cases outside ordinary discourse - experimental laboratory tests, for instance - to suppose that some sort of explicit rule-application is going on, but then the interesting question, which remains unsatisfactorily addressed, is how far one's explanations of these cases extend to others, and why.

Considerations such as these seem to shed doubt on remarks such as Johnson-Laird's confident assertion that 'practical inference' is likely to involve large numbers of rules of inference (as in natural deduction). Indeed, its aim can be described as inferential, but it's

unclear that reflection on any sort of formal system will provide much illumination on how it achieves it. An unfortunate aspect of the matter, as I have been suggesting, is that comparison with these systems is inclined to lead to the sometimes inappropriate assimilation of psychologically heterogeneous phenomena as 'inferences'. Significantly, Johnson-Laird's later account is explicit in its denial that a rule-system is needed for any kind of inference. In my view, however, he still gets into trouble by failing to take note of the large range of different kinds of inferential performances that exists. To the extent that he makes his theory precise - as when he deals with syllogistic inference - I suspect that he limits its application; he is over-ready to assume that *all* circumstances in which one can see a reasoner as solving a syllogism, will succumb to his mechanism. We shall find more arguments to this effect in Part II.

2.3 Reasoning and Behaviour

We now find ourselves returning to the question of what, then, 'reasoning' is. It needs to be distinguished, I have claimed, from a vast mass of processes which might be called 'inference' in various forms, but whose explanation very likely belongs in a different domain. Has the person in (3) *reasoned*? There is little doubt that the typical answer would be 'yes'. It seems to me, though, that this is a mistake, not so much because (3) is clearly *not* a case of reasoning, but because the notion of reasoning needs to be clarified before we can tell; it needs to become theory-relative; we need to elaborate a theory of reasoning based on clearer cases (eg. laboratory tests) and then ask whether it extends to ordinary discourse. We might find out that it does, or that if we adapt the theory we can make it cover both kinds of cases, or that no such adaptation is feasible, and that cases like (3) are best regarded as cases simply of language use, describable as having involved an inference, but not suitable for description as cases of reasoning because not demonstrably involving what we have come, or decided, to call 'reasoning processes'. The error to avoid is supposing that an external description of a piece of behaviour in itself provides clues as to the nature of the processes responsible for producing that behaviour.

If one asks why we should not start from ordinary discourse and see whether a theory of that will extend to laboratory cases, the answer should be that this is in principle an equally suitable course to take. In practice, however, it is obvious enough that it makes one's initial task much more difficult. For one thing, it's not at all obvious that 'ordinary discourse' describes something amenable to treatment as a single phenomenon. There is a clear methodological advantange in being able to create a relatively simple and testable theory to start with and then investigate its ramifications. The other approach is likely to lead to unproductive floundering in a sea of complex and unstructured data.

So we return also to the suggestion advanced before this rather extended digression, that whether a performance has involved reasoning depends on the 'level' at which processing took place in producing it. Unfortunately, we seem to lack any accepted empirical criterion for deciding the level at which processing has occurred (as well as an agreed criterion for when it counts as reasoning). Intuitively, to steal Johnson-Laird's concept, reasoning has to involve the drawing of several connected psychologically basic inferences - but how do you identify such cases? One approach (we have already seen Johnson-Laird mention it) might be to use introspection: we ask for instance the snooker player whether he reasoned about such-and-such an aspect of his problem, and accept his judgment. (More naturally, one might ask if he *thought* about it, and then on the present account one would clearly be left with work to do in finding out just how he had thought.)

Introspection is commonly treated with a good deal of suspicion, and rightly so. If one comes to be able to report, instrospectively (or retrospectively) how one has done something which one has hitherto done quite automatically, the nature of the report depends on how one *conceptualises*, in some sense, the performance in question, and this can depend on all sorts of external factors, such as how one was taught, etc. There are two points at which worry enters here: for one thing, there is no way to tell whether the report is accurate, rather than some kind of confabulation; and for the other there is no *prima facie* reason to suppose that one can gain introspective ability with respect to some performance, without changing the processes whereby it is produced. But these may be problems mainly where the concern is with using introspection to test the correctness of a hypothesis about the nature of processing. The present idea is merely to elicit from the subject whether he is aware of any processes at all having taken place, and this perhaps can be done in a way which does not 'lead' him by suggesting any particular conceptualisation of those processes; but I have to grant that to some extent we venture here into unfounded speculation.

Even if people know they have done something, of course, they may not know *how* they have done it. Dennett (1978 165) recycles a remark of Lashley's, to the effect that all we are ever aware of, in mental processing, is the *result*. All the same, one can commonly decompose a case of processing into subprocesses, for each of which a result is available. At some level, the thing is a set of connected 'black boxes', but sometimes one can say what boxes have been connected - and sometimes one cannot. It was like this with the snooker player: we wanted to say that he was (quite probably) aware of intermediate results in the process of deciding which ball to pot next, and where to leave the cue ball, but that (quite probably) most of the decisions about how to bring this about were to him automatic, and he would be unable to say anything about the stages whereby they occurred; he simply knew and acted upon the results. The decisions of the first sort were those we felt most inclined to describe as cases of 'reasoning'. If the reasoner can't say

what black boxes he connected, he effectively had for that (sub)process only one unanalysed black box; he had only a result, and it becomes difficult to describe what he did as a process of reasoning. (Except, perhaps, as a single psychologically basic inference - there's an obvious relationship between my proposed interpretation of this term, and my usage of the notion of a 'black box'.)

A likely complaint here is that this is a ludicrously tight, and at the same time vague, restriction on what counts as reasoning. Apart from anything else, there is a strong tendency to regard the description of a task, for instance, as defining whether or not it is a reasoning task. If people are presented with cards bearing two premises, from which they are asked to draw a conclusion, it seems hard to deny that if they comply they *must have reasoned*, no matter what they can or cannot say about how they carried out the task, at any rate so long as their responses are plausible. When a subject says his correct answers 'just came to him', we are surely almost bound to suppose that he *reasoned unconsciously* - which does not appear to be an internally incoherent notion. Moreover, it will be asked, which among the variously confused protocols that one gets from subjects one questions about these things, are to be taken as an account of subprocesses in a reasoning performance?

At this point, a danger arises that we will conflate what are really (at least) two distinct difficulties in looking at reasoning performances. On the one hand, we have the problem that we have been considering here - the question about the level at which some task has been done. On the other, we have a problem well known to the literature in this area, which is the debate over the matter of whether people carrying out these tasks are or are not rationally following some course of logical deduction. My claim that these are distinct depends on the view that whether the processes we are interested in accounting for are 'logical' or not, is something that can be considered separately from the question as to whether any such processes have occurred. The danger of confusion arises from the nature of certain theories, typically supporting the 'illogical' picture of reasoning, which propose, eg, that subjects apparently introspecting a rational process are in fact confabulating a plausible explanation for behaviour they cannot otherwise justify, and of whose real aetiology they are quite unaware (*cf.* the Wason/Evans 'dual-process' theory - Evans 1982 ch. 12.). We have a problem here in trying to support the idea that subjects can be aware of having done something, while being perhaps so *completely* unaware of how they did it. At best, on this type of theory, the subject can say that black boxes were connected, but he has no actual clue as to which or how. We now have no real access to intermediate results, except for the impression that there were some! And how can we tell that even this impression is not erroneous?

* * *

We are arriving now at a position for reasoning which resembles some of those copiously discussed in the philosophical literature on dreaming. How can we tell (a) whether and (b) about what someone has dreamt? The obvious answer is to ask him. It's certainly hard to imagine answering (b) in any other way. But the suggestion is made that (a) could be answered on the basis, say, of EEG output, examination of eye movements, and the like. A theory in vogue at one time held that dream reports were constructed in response to traces left in the brain by some almost instantaneous process that certainly could not have been an experience at all like that reported; they were *post hoc* confabulations. A commoner view nowadays is that dreaming occurs during particular periods of sleep, particularly in association with rapid eye movements, and it is accordingly more likely to be experienced as described. Both of these theories, however, can in my view be seen to be logically dependent on the strength of the report. While it might be possible in some given case to cast doubt on a subject's claim to have dreamt - perhaps even, in view, say, of the duration of certain observations, to question the plausibility of his account of the course of the dream - it would make no sense to do this at all generally, because it would undermine the foundations of supposing that the reports and the other evidence related to the same phenomenon in the first place. The possibility of such questioning is based on the assumed reliability of a correlation between a great many dream reports and other kinds of external observations, and the distinctive discrepancy of a given case. If one is going to study dreams, one simply has to suppose that people's reports of dreams are in normal cases reasonably accurate, or else be prepared to concede that 'dreams' may not after all be what one is studying. One might end up holding that there is no single phenomenon that can be definitely pointed to as the normal cause of the propensity to report dreams, in which case so much the worse for a theory of dreams (*cf.* Dennett's view about pain - 1978, ch. 11).

Now, perhaps it's like this with reasoning, as well. Unless the notion can be given some wholly behavioural and circumstantial definition, so that for instance the production of valid conclusions to premises, perhaps with particular response latencies, etc., is *ipso facto* the performance of reasoning - which surely is both implausible and anathema to cognitivists - then this seems inevitable. People have done something that is at least a candidate for a case of reasoning, if they say they have.

It would never do anyway, of course, to find ourselves sustaining the claim that a *task* can be essentially a reasoning task. A task may be describable as a reasoning task, but it can also always be described in some other way. People (eg. who have done it often) might be doing simple memory retrieval, or guessing at random; and surely we are not going to call *these* approaches 'reasoning'. This is not to prejudge the issues alluded to above, which we shall presently look at in more detail, by supposing that reasoning has to be describable in terms of processes following a strict path of logical inference. But

there is a strong intuition in favour of the view that people have typically to be able to offer a sincere report of how they arrived at the conclusion, whether it appears logical or illogical, and however misguided our theory may later imply it to be. In some sense, what is important is *the subject's* description of a task (not the experimenter's), but only the subject can tell us what this is. That is to say, people need to be aware *that* they arrived at the conclusion, even if they are wrong about *how*: they need to have *understood* the nature of the task, to have described it to themselves in something like the experimenter's terms, or at least in terms recognisable as descriptive of some reasoning task. It needs to be stressed that this is not a firm precondition of any *case* of reasoning (even as people may dream without being able to report the fact); but that it is typical is a presupposition of any *theory* of reasoning.

A theory of reasoning, then, should begin as a theory of performances describable in behavioural terms as the drawing of inferences, typically in circumstances where the subject was aware of actually doing the task (not simply of having done it) - can say something about how the task was done. In my opinion, even Johnson-Laird's 'test case' for cognitive theorising, quoted near the beginning of this chapter, is vulnerable to this point: it is highly doubtful, in default of further evidence, that anything like the processing of a syllogistic argument was performed, and possibly even that the subject of the story reasoned at all.

2.4 Reflective and Unreflective Performances

Earlier, we mentioned a worry, aside from the worry about whether the report is unwitting confabulation, about whether the ability thus to report on one's processing at all might change that processing. By this is meant the following: if someone is prodded about how he is doing some task, until he does it *reflectively* - which is to say in some sense paying attention to what he is doing at the same time as doing it - how do we know that this has not lead to a change in the processing going on as compared to his previous unreflective performance? The answer to this, it seems to me, is that we cannot know such a thing. There is in fact no reason to suppose that the processing has *not* changed, in order for this apparent access (even if it is illusory) to have become possible.

In general, it seems altogether likely that tasks done reflectively, on the one hand, and unreflectively, on the other, are done somehow differently in each case. We familiarly find variations in performance between the two cases, on otherwise similar tasks, so that people sometimes come to feel, when they reflect, that their unreflective conclusion was incorrect. In fact, a common state of affairs in cases of the latter kind is that subsequent unreflective performances are unchanged, even though supposed to be wrong. Changing them can require a substantial educating effort. If I find myself persistently in error over, say, some

particular class of mental arithmetic calculations, I may practise a method for solving such problems reliably until I always get them right, eventually without reflecting on the method, whenever they occur. But now, how can I tell whether the method I am using is *still* the one I practised? How could anyone tell? Have I, in 'internalising' the method, changed it importantly (or even completely)? I can no longer report on it by recollection; I may believe that I know how I'm doing it, but I no longer claim to be able to remember what I did in a given case.

What is true in this example holds *a fortiori* in circumstances where we have no evident link at all between a reflective and unreflective performance, save a similarity in the logical form of descriptions of them. What now might be thought to have happened, though, is that we have left no room for the possibility of theorising about unreflective reasoning. We have not, but we have drawn attention to the fact that such a theory is going to be parasitic on a theory of reflective performances, because it is only in the latter cases that we can get a grip on the phenomenon we are intending to study. We have also indicated that considerable care needs to be taken in extrapolating from reflection. Unreflective reasoning can be identified only by the gross features of the behaviour and the circumstances, only by our making an assumption that some process of a sort we might want to call 'reasoning' has occurred; and our very notion of what such a process is depends on our theory of reflective performance. We hope that our assumption can be checked by constructing a theory predicting additional phenomena such as behavioural proclivities in cases yet untried, response latencies, etc.

We want to say that certain processes would not count as reasoning and an attempt at an account of these follows almost immediately. This remains all rather vague and pretheoretical, of course, but nonetheless evidently based on the experience of reasoning, and on accounts drawn from others' reflective solutions. The point is that making it more precise and theoretical involves firming up these intuitions, without altering their origin. Our theory of reflective reasoning is going to be crucially important to any theory of unreflective reasoning *qua* reasoning; and in fact it's a mainstay of the argument in the first part of this thesis that we could not provide any truly *cognitive* account of these performances without such appeal, because to do this requires the interpretation of proposed processes in intentional terms, and these are irrevocably rooted in our experientially-based cultural notions of what we are up to when behaving in certain ways (what is frequently known as 'folk psychology').

Notice that this is not an attempt to involve phenomenology, or anything of that kind; in fact I am sympathetic in many ways to the view of self-consciousness which takes it to be an essentially theoretical construct based ultimately on observations of behaviour and various devices aimed at its explanation (*cf.* Churchland 1979, Dennett 1982a,b, Evans

1982, Nisbett and Wilson 1977, but see also Egan 1983). An important determiner in many cases, though, is specifically *verbal* behaviour, and the kinds of it that occur or can be elicited in conjunction with non-verbal behaviour. On any sensible version of this theory, it is in fact only language that allows us to arrive at self-consciousness or introspection. In a very Wittgensteinian fashion, we can regard public, linguistic interactions as providing the basis for constructing our notions of a private, mental life. But then it can be seen all the more clearly that any effort to reconstruct mental processes will have to be firmly rooted in the publicised world of what we (and others) can say we have done.

2.5 Reasoning Processes

It seems appropriate now to return to a question fenced around a couple of times already, which is the question about what kinds of processes - given a presumably accurate description of them - count as reasoning processes. It's evident that not just *any* kind of process is going to count as reasoning. In tasks where subjects are offered premises and a set of possible conclusions to choose from, there is a good chance that they may do no reasoning at all, in many cases, when selecting the conclusion. They may even guess at random, if they get bored. The likelihood that reasoning has been used can be increased by presenting only the premises: at least if people's conclusions are plausible, here, they are likely to have used something more than simple guesswork. No amount of task engineering of this kind is going to *guarantee* that the performance has involved reasoning, however. Even should the responses given be always *correct* (in some understood sense), uncertainty may remain.

Reasoning, I think we want to say, is closely bound up with the notion of *argument*, and thereby that of *proof*. To reason one's way to a conclusion, is to produce (however implicitly or internally), an argument for that conclusion. It is to come to believe that one can prove, demonstrate the truth of the conclusion, from the given premises. To reason correctly is therefore naturally to produce a valid argument. However, this characterisation tells us really very little, since there are a number of different understandings of the ideas of proof and argument.

The intuition described here is similar to one of Johnson-Laird's, expressed in the slogan that subjects, properly to be said to have reasoned, have to be seen as having reached their conclusion *for the right reasons*, not 'as a result of processes that do not suffice to establish validity'. How this is interpreted by Johnson-Laird is something to be considered later, but it clearly captures the idea that people who arrive even at systematically correct conclusions, have not reasoned unless their processes have been of a certain sort.

Suppose we try to work backwards from cases where processing seems clearly *not* to be reasoning. Simple memory lookup and random guesswork are obvious examples, but what about these makes them non-reasoning processes? We can note that in the first case, information is sought from another source (memory), while in the second the information in the premises is not used, or not used very much. We might therefore suggest that in reasoning, the information in the premises is used, and used exclusively, to arrive at the conclusion. This entails that all sets of premises will be treated in the same way by these processes. Given standard accounts of what deductive logic is, this is an unsurprising suggestion: the conclusion contains nothing that was not given in the premises, and is derived from the premises by some reliable means. Notice, though, that there are no implications here about the nature of the means, other than that it is reliable and general. Some sort of generate-and-test system might very well qualify, since we can see that while the initial generation may be even quite random, the subsequent stages will have to employ information from the premises again, in order to weed out unwanted conclusions.

Taken in the large, it looks as though this will be too strict. Surely, we have to allow that reasoning can bring in 'world-knowledge', and certainly specific verbal knowledge? But these requirements can be accommodated by regarding the additional information as additional premises. It is always allowable to introduce extra premises, so long as they are then treated in the same way as the original ones; the crucial point is that the processing remains *general* and not bound to specific cases. There seems to have to be a mechanism such that arbitrary premises can be inserted at one end, and something appropriate result at the other.

Since 'something appropriate' is actually (in the ideal case) valid conclusions, this amounts to a proof system for some logic. We shall see in chapter 4 that in the context of computationally based theories of reasoning we have here a characterisation of the subject-matter that becomes problematic to sustain. Even before we get to that point, though, there's something uncomfortable about it. Surely psychology is an empirical science, the aim of which is to describe just what goes on inside people, what causes their behaviour, regardless of whether that behaviour has another characterisation as, eg, the drawing of *valid* inferences. Psychology, one feels, should not take the view that people's conclusions *ought to be* valid; all it seeks is an explanation of why and how they were drawn.

I think what emerges here is that if reasoning is defined in this way, then we know that (and exactly how) it occurs reflectively, but we don't know that it *ever* occurs unreflectively. If it does, the fact will be very difficult, perhaps impossible, to demonstrate. In our present state of knowledge, at least, it is quite conceivable that unreflective 'reasoning' performances such as the ones in Johnson-Laird's test case are all manifestations of some complex, commonly accurate, but nonetheless far from guaranteed system of

guesswork, based on all kinds of heuristic applications of specific background facts about the topics mentioned in the premises. It is even plausible, especially on an evolutionary view: why should Nature take the sledgehammer of abstract logic to the relatively small nut of ordinary context-bound problem-solving?

Such a state of affairs would of course reduce logical form to the status of one, perhaps minor influence, among very many, on the outcome of the processing. And this is important, from the point of view of creating a theory along algorithmic lines, intended to be at all general. I want to suggest that should our cognitive nature actually be like this, the prospect of creating a coherent cognitive theory of reasoning vanishes. At best, we can hope for a number of disparate theories covering performances in different domains. The following chapters will explain in more detail the reasons for supposing so, as likewise for supposing that the ideas of competence and rationality will lose their grip in discussions of performances of these kinds. One might suggest that we have left room for *reflective* performances to be characterised by a logic-based theory; this relates to discussions in Part II, but my answer to it turns essentially on the point that such a theory will tell us more about the domain of logic than about the processing involved, for I suspect that any remotely adequate processing account of reflective phenomena will be so complex as to obscure the logical aspects fairly thoroughly.

2.6 Conclusion

What we have really tried to do in this chapter is uncover some dubious practices in current cognitive psychological theorising about reasoning, and demonstrate that further thought about the question is actually needed. We have tried also to make suggestions about the direction work ought to go in. Our purpose is not simply to browbeat psychologists for ignoring some important methodological groundwork required by their enterprise, but also to draw attention to issues which are important in the argument of the following chapters. In particular, it will be henceforth supposed that a theory of reasoning is a theory directed in the first instance at some range of clearly identifiable tasks, rather than at discourse phenomena in general.

3

Cognitive Models

The point of this chapter is to expose certain underlying assumptions of a view of cognition commonly accepted in cognitive psychology, and cognitive science. There is nothing particularly original in this exposition, and it is not intended that it should involve either criticism or defence of the assumptions being investigated. I wish merely to establish certain premises which will be useful in the subsequent argument, and many important issues are simply sidestepped. It is not necessarily the case that all cognitive scientists would accept the assumptions laid out here, at least in the form in which they appear, but the idea is that these assumptions are sufficiently widely accepted, in some form or other, to characterise an approach that it is worthwhile to examine.

3.1 Models as Algorithms.

The general view of the mind presupposed in these areas of study (at least by the majority of participants) is that it is in some sense a *computational* device or mechanism. In trying briefly to establish why this is and what it entails, I shall draw heavily on the account offered by Pylyshyn (1980). It will be claimed that cognitive models depend on the postulation of *rules* whereby the mind is to be seen as operating, and indeed that these specify processes of an *algorithmic* nature, which are assumed mentally to occur. (It should be noted here that there are various popular notions in cognitive science, particularly some of those falling under the umbrella of 'connectionism', that fit this description only doubtfully, but it is beyond the scope of the present discussion to investigate the consequences of this.) It will be urged, further, that the postulation and, more particularly, the interpretation of these rules, depends on the assumption that the behaviour being modelled is rational. This point is rather significant in view of the arguments to come in later chapters, but it will be treated here only in an introductory fashion. Drawn substantially from the views of Dennett, it is probably somewhat more controversial than much else in the chapter.

The object of cognitive psychology is to capture, and then explain, certain regularities in human behaviour, which emerge under a particular kind of description. The kind of description in question is, essentially, one which relates the organism to its environment, and allows us to state such relationships in illuminating ways. This is familiar, of course, in such constructions as: X desired that *p*, and believed that *q*, and that's why he did *A*; where *p*, and *q*, are each descriptions of some state of affairs either in or previously or potentially in the organism's (or person's) immediate environment, and *A* is a description of X's behaviour as an *action*, not just movements, etc. In pretheoretic 'folk-psychology', of course, we routinely describe and account for people's (and animals', etc.) behaviour in just these terms of belief, desire and action. We say, to steal Pylyshyn's example, that someone is *running out of a building* because of his *belief that it is on fire*, and his *desire to escape the fire*. We can account for the acquisition of the belief in various ways - the smelling of smoke, the recieving of a 'phone call - and hence an important regularity can be stamped upon a number of events quite diverse under other (eg, physical) descriptions. What is essential in all this is that the behaviour of the organism (person) can be *interpreted* as a series of actions, directed toward some goal, and informed by a description of the environment of the action (this is particularly noticeable when anomalous behaviour is explained via the postulation of *false* beliefs about the environment). Also essential is that beliefs and desires are not just arbitrarily combined to explain action: there are severe logical constraints, which one might regard as *rules* governing the procedure of constructing explanation-descriptions of this kind. (We shall discuss these later.)

The cognitive theorist is disposed to think of these informational and goal states as being (internally) *represented* by the organism, these representations then being manipulated according to certain rules (essentially, serving as premises in a practical deduction) in a process leading to the determination of action (*cf*, in particular, Fodor 1975). As Pylyshyn says (*op. cit.* 112), we have here generalisations which 'can only be stated in terms of the agent's *internal representation* of the situation (ie, in mentalistic terms)'. (It cannot go without comment that thus to equate mentalism with representationalism, assuming that the 'mentalistic terms' are straightforwardly what is represented, is a somewhat controversial assertion, unargued by Pylyshyn; but it is commonplace in the field. We shall have cause to note this point again below.)

Given this characterisation of the phenomena to be explained, it is natural to turn to computation for an explanatory theory. For, as Pylyshyn notes,

> . . . computation is the only worked-out view of *process* that is both compatible with a materialist view of how a process is realized, and that attributes the behavior of the process to the operation of rules upon representations. In other words, what makes it possible to view computation and cognition as processes of fundamentally the same type is the fact that both are physically realized and

both are governed by rules and representations.

The point to be focussed on here is that computation is not intended to be seen merely as some sort of useful *model* of cognition, nor is it just that there is a useful *analogy* between the two: Pylyshyn's (and, in general, cognitive science's) claim is that cognition literally *is* computation - in some sense. Naturally, this is not supposed to involve regarding the brain as constructed in literally the same fashion as contemporary electronic devices, nor that it functions precisely similarly. The notion of computation is complex, and there are no doubt possible kinds of computational device yet to be produced, or conceived of. However, there is supposed to be this feature central to *all* conceivable computational processes, that they 'are governed by rules and representations'. It is true that this is stipulative in that, eg, it simply rules out 'analog' devices as not being properly computational, since they don't seem to involve rules defined over representations, but perhaps Pylyshyn is justified in holding that it is not arbitrary. Support is garnered from Brian Smith (1982 - see below) and Haugeland (1978), at least if Pylyshyn will agree that by 'computation', he means 'symbolic computation'.

Mental activity, then, is to be regarded as 'the execution of algorithms'; and the construction of a *cognitive model* - an attempt at explaining some feature of this activity - must accordingly involve the attempt to specify an algorithm (or 'program'). It should be stressed, somewhat incidentally, that this has nothing to do with so-called 'non-algorithmic', 'heuristic' programming techniques, eg. in chess-playing programs: even here, the operation of the heuristics must be in accordance with some algorithm, if it can be executed on a computer. It has been claimed (eg. Dreyfus 1972) that human performance on at least some tasks requires *unprogrammable* heuristics; but this remains to be shown, and if it is shown to hold on a wide scale, then the whole project of computational cognitive psychology will possibly have been a mistake. (In the conclusion of this thesis, however, we will recommend a position which entails the partial acceptance of a position related to this.)

One must be careful to say how much is involved in specifying the algorithm. We need, of course, more than just an 'input-output' description of the *'behaviour'* it is supposed to produce, since many different algorithms could be equivalent in this respect. The 'rules and representations' must both be described in detail, so that one can distinguish between different algorithms, which might be proposed to account for the same phenomena. Pylyshyn adduces a notion of 'strong equivalence' of algorithms, which unfortunately has to be left informal and intuitive; but the suggestion is that it can be approximated by some other notion, to which end we are offered the following:

> . . . all processes that, for each input, produce 1) the same output, and 2) the same measure of computational complexity, as assessed by some independent means, are referred to as complexity-equivalent (117).

The idea here is that complexity-equivalence can be empirically investigated if the 'independent means' of assessment is empirical; and it is held that reaction-time data, carefully interpreted, might be such a means. Hence, we require algorithms to be specified such that their properties with respect to complexity can be judged and tested, since the goal of the research is ostensibly to discover the algorithm that the subject actually uses.

We speak of the algorithm as consisting of rules defined over representations, but the consequences of this have to be recognised. In a physical realisation of the process, its actual state-transitions are caused physically, and to that extent depend wholly on physical laws. But we have an account of these state-changes in terms of the properties of the domain of objects (whatever it is) appealed to in an interpretation of the representations. For instance, if the program carries out addition,

> . . . in order to explain why the machine prints out the symbol '5' when it is provided with the expression '(PLUS 2 3)', we must refer to the meaning of the symbols in the domain of numbers. The explanation of why the particular number '5' is printed out then follows from these semantic definitions (ie, it prints out '5' because that symbol represents the number five, 'PLUS' represents the addition operator applied to the referents of the other two symbols, etc., and five is indeed the sum of two and three (113).

This works only because all the representations are somehow physically distinct, and in such a way that when they interact with the physical realisation of the rules, the appropriate result emerges. The 'realisation of the rules' is known as an 'interpreter', because of this property; and its general properties are known as the 'functional architecture' of a 'virtual machine' - in this case, an adding machine. It should be noted that the description of this is always in purely functional terms - hence 'virtual' - and it could be implemented in principle in any number of ways, even as a program for some other machine; Pylyshyn, though, always assumes the 'cognitive' architecture to be implemented directly in the biological or neurological substrate.

A consequence of these observations is that all the relevant semantic distinctions which are captured in the representations used, have to be realised as physical (one usually says, 'syntactic') distinctions between these representations; and all these must be correlated with functional differences in the architecture of the interpreter:

> This is what we mean when we say that a device *represents* something. Simply put, all and only syntactically encoded aspects of the represented domain can affect the way a process behaves (113-4)

The same principle is also visible in Fodor's famous 'formality condition' on computational representation (Fodor 1980).

The discussion here is remarkably similar in many respects to the account of 'computational process' offered by Brian Smith (*op. cit.* 23-27). A process is understood, initially,

to be somehow defined only in terms of its 'surface', ie. its behaviour. Smith adduces the notion of a 'reduction' of a process, whereby it is broken down and seen to consist of, basically, an inner process operating over some 'structural field' of semantically interpreted symbols. According to Smith, it is essential to a process' being computational, that it is susceptible to such a reduction. Clearly enough, the inner process here corresponds to the notion of an interpreter, while the field of symbols amounts to an inscription, in some appropriate notation, of the algorithm in question. In principle, there can be many embedded reductions, but, significantly, the inner process(or) of the last reduction is only ever described in terms of its surface or behaviour, and thus is not revealed to be computational. This corresponds to Pylyshyn's view of the computationally opaque nature of the functional architecture of a given machine.

What seems completely clear is that on this account a major task of cognitive psychology must be to elaborate as far as possible what Pylyshyn calls 'the functional architecture of the cognitive virtual machine', since this and the range of possible algorithms (up to strong equivalence) that it can execute are strongly mutually determining, like a lock and key: '. . . different architectures are in general not capable of executing strongly equivalent algorithms' (124). The main point for us here is that the functional architecture decides the *primitive* functions that the algorithm can appeal to. An architecture might be such, for instance, as to offer *addition* as a primitive, but not *multiplication*, in which case the latter would have to be implemented perhaps as repeated addition; if, though, multiplication *were* primitive, then we would not have to (indeed could not) offer an account of it in terms of an algorithm interpreted by this machine. Pylyshyn's contention, here, is that such primitive operations are not part of the subject-matter of *cognitive* psychology, for which they must remain fundamentally unanalysed. It is not that these primitives of the cognitive architecture have to be left utterly mysterious, but rather that any account of them must be *non-cognitive*, perhaps neurophysiological, or in terms of analog processes of some kind. Those, such as Dreyfus and the followers of Gibson, who seem to urge that everything in the brain is analog, are thus seen to be urging that *everything is primitive*, and thus that there is *no scope at all* for cognitive explanations. Cognitive psychology is clearly committed to the assumption, which as Pylyshyn stresses ought to be regarded as empirical, that this is false.

Pylyshyn, in fact, goes on to establish 'algorithmic accountability' as definitive of the now technical notion of a cognitive phenomenon, and suggests that the conditions under which it is manifest can be identified with those in which a phenomenon is or can be systematically influenced by cognitive factors such as 'changes in instructions or the information-bearing aspects of the context'. The apparent circularity here is explained away by supposing that what is going on is a process of gradual evolution and as it were theoreticising of our basic intuitions about what cognitive phenomena are. Thus is

Pylyshyn led to produce his well-known condition that primitive functions and properties have to be *cognitively impenetrable*, by which he means that they cannot be directly influenced by any factor which has a cognitive description (155).

Pylyshyn's view here is, perhaps, a little extreme. Haugeland (1978 *op. cit.*), for instance, holds that there is, or might be, a hierarchy of 'intentional black boxes', several of them interacting at any given level, but each of them decomposable into a number of other interacting black boxes. This is rather similar to Lycan's (1981) and Dennett's (1978) 'homuncular functionalism', and might lead to the suggestion (*cf.* Haugeland's commentary in Pylyshyn *op. cit.* 138) that the cognitive functional architecture ought to be seen as just another level of computational processing, itself reducible after Smith's fashion to a further processor and a different algorithm. The main value of this would be its admitting an idea that Pylyshyn does not discuss, which is that of *compilation*. In conventional computers, algorithms are often not executed directly, but instead translated (compiled) into a possibly different but behaviourally equivalent algorithm, which is executed instead. If one has an interpreter for a high-level language implemented in some lower-level language ('machine code', perhaps), one can sometimes selectively compile just certain routines or procedures, components of some larger program, into the lower level language. They then, from the point of view of the top level, *become primitive*. The possible application is that skills, for instance, which at first can be practised only with careful thought, and attention to every step, might eventually become compiled and thereafter operate very much more quickly, and without the possibility of cognitive penetration. Such a notion might turn out to be of great psychological value, although Pylyshyn seems obscurely to doubt this (165).

A significant further matter is that the actual, physical instantiation of the functional architecture might at certain points become of interest for the cognitive theorist. He will have to consider cases where the execution of the algorithm fails to proceed smoothly, for some reason or other. This might be due, for instance, to arbitrary limitations on memory-space, or the appearance of some kind of interference in certain situations. Precisely what circumstances these phenomena occur in, may well be a useful pointer to the nature of the architecture.

3.2 Representation and Interpretation.

The task of the cognitive-model builder, then, is to propose an algorithm (or set of algorithms) whereby the behaviour which interests him might be produced. This has to be supported by arguments showing that it is the *right* algorithm, the algorithm actually producing the behaviour of his subjects. So far, however, this is a purely *formal* enterprise. We have seen that the machine executing an algorithm has no access to the domain of

semantic interpretation of its representations, save insofar as this is mirrored in syntactic features and distinctions. In principle, then, one could carry out the project of proposing these algorithms without ever mentioning the interpretational domain, describing one's results merely in terms of their syntax. Indeed, Stephen Stich, in his 'Peer Commentary' on Pylyshyn's paper, suggests that semantics is just irrelevant to cognitive psychology; talk of *representation*, Stich says, 'is simply excess baggage' (152). This is a topic addressed by various people - eg. Fodor (1980) - and enlarged upon at length by Stich in his later book (1983). It would delay us indefinitely to go into it here. Pylyshyn rejects Stich's thesis because, he holds, it would be to 'abandon the goal of relating a computational psychology to a folk belief-desire psychology' (161), and to do this would be to lose the explanatory powers of the theory with respect to the behavioural regularities mentioned at the outset. The point is that these regularities, as noted above, are stated under an interpretation of the behaviour, and cannot be satisfactorily explained under that interpretation by a wholly uninterpreted algorithm. (We want to know why the computer prints out *five*, not merely '5'). It would have to appear ad hoc; its features could not be motivated in the appropriate ways; it would lack the required relationship with the agent's *environment*. As Pylyshyn says:

> The only way both to capture the important underlying generalisations . . . *and* to see . . . behaviour as being rationally related to certain conditions, is to take the bold but hightly motivated step of interpreting the expressions in the theory as goals and beliefs (*loc. cit.*).

This is a revealing remark (even if overstated: as will appear, it is questionable whether the expressions are to be interpreted *as* beliefs and desires), and it prompts attention to the second main assumption of cognitive theorising, which I proposed to introduce in this chapter, *viz.* the assumption of ***rationality*** in the agent.

Discussions of this assumption are not uncommon in the literature of philosophical psychology, particularly since Quine and Davidson. In those contexts, it often takes the form of a 'principle of charity' in radical translation: if one fails to assume that the native is responding rationally to his environment and its stimuli, then translation cannot get off the ground. We shall have cause to examine this claim in later chapters, but problems of translation are, naturally enough, tightly related to problems of folk-psychology generally. As Grice (1975) has made as clear as anyone, one's interpretation of a speaker's utterances depends on one's views about his psychological states, in particular his propositional attitudes; his intentions, or beliefs and desires. In general, utterances, particularly regarded as speech-acts, are just a subclass of actions, and the contention is that the whole enterprise of interpreting behaviour in terms of actions, verbal or otherwise, depends on the rationality assumption.

A particularly clear and forceful statement of these points, which has the added virtue of relating them specifically to cognitive theorising, is provided by Daniel Dennett (1981a; see also 1981b). He contends that any kind of 'intentional' psychology - a psychology, that is, which depends on the attribution to organisms of contentful, or in some sense information-bearing, states - has to begin by recognising the organism as an instance of what he calls an 'intentional system'. What this entails, Dennett explains, is that the behaviour of the organism can be usefully predicted by applying to it such notions (in a suitably systematised and clarified form) as 'belief', 'desire' and 'action'. There is a sort of 'abstract calculus' according to which one can manipulate these attributions in order to arrive at one's predictions. Dennett holds that this kind of approach depends not at all on any kind of knowledge about the internal constitution of the organism in question, nor on any notion of the (physical-causal) aetiology of the observed behaviour, just so long as that behaviour exhibits the right sort of 'patterns'.

What we have here is a version of an overtly instrumentalistic 'logical behaviourist' doctrine, differing, eg, from that of Ryle (1949) in that all intentional attributions are inter-linked, and not accountable in isolation. (Which avoids the major thrusts of many of Ryle's critics, particularly Fodor, 1975.) Dennett regards his intentional systems theory as a kind of distillation of much that is important, and effective, in folk-psychology; the rest, he believes - whatever can be said about the actual realisation of intentional systems in, for instance, human beings - gives rise to a discipline he entitles 'sub-personal cognitive psychology' (SPCP).

The success of the intentional strategy in predicting behaviour depends on that behaviour being rational. The attribution of intentional states to any system depends on the assumption that the system is rational, or *well designed* (which Dennett supposes is what evolution selects for). These remarks amount to declaring an equation between being rational and responding to 'intentional characterisation according to the rules of attribution'; these rules are in some sense definitve of rationality, and vice versa. Dennett insists that intentional theorising, and indeed folk-psychology itself, are *normative:*

> A system's beliefs are those it *ought to have*, given its perceptual capacities, its epistemic needs, and its biography . . . A system's desires are those it *ought to have*, given its biological needs and the most practicable means of satisfying them . . . A system's behaviour will consist of those acts that *it would be rational* for an agent with those belief and desires to perform (1981a 42-3).

This is, of course, an idealisation, since no system is ideally 'ensconced in its environmental niche', and no (at least, no human) system is ideally rational. When these things fail, 'special stories' have to be told; but the apparatus works just to the extent that agents approximate the ideal.

If folk-psychology and intentional systems theory depend so crucially on the rationality assumption, then no less so, according to Dennett, does SPCP. This is simply because, where semantics is concerned, the former doctrines are prior to the latter. Dennett's view of SPCP entails that its main task is to account for the brain's functioning as a *'semantic engine'*, when really all it can be is a *'syntactic engine'*. This parallels the remarks above about computational systems being essentially purely formal. The brain, evidently, has to

> *mimic* the behaviour of the impossible object (the semantic engine) by capitalizing on close (close enough) fortuitous correspondences between structural regularities - of the environment and of its own internal states and operations - and semantic types (54)

- which recapitulates the point about semantically relevant distinctions having to be mirrored by syntactic ones. What one necessarily ends up with, then, are

> . . . systems that *seem* to discriminate meanings by actually discriminating things (tokens of no doubt wildly disjunctive types) that co-vary reliably with meanings (55).

I want to maintain that SPCP as envisaged by Dennett is essentially the same discipline of cognitive-model construction that was discussed with reference to Pylyshyn. This seems uncontroversial enough, given Dennett's assertion that

> It is the task of sub-personal cognitive psychology to propose and test models of such activity - of pattern recognition or stimulus generalization, concept-learning, expectation, learning, goal-directed behaviour, problem-solving - that not only produce a simulacrum of genuine content-sensitivity, but that do this in ways demonstrably like the ways people's brains do it, exhibiting the same powers and the same vulnerability to deception, overload and confusion (55).

Dennett denies that we are likely to find here anything corresponding to the 'personal-level' categories of belief and desire; *contra*, eg, Fodor (1975), he thinks it implausible that propositional attitudes as such are couched directly in some language of thought, but rather that they are 'emergent' regularities exhibited by the algorithms that actually are operative (*cf.* Dennett 1978, esp. chs. 2, 6). It is not, as Pylyshyn has been sniped at above for suggesting, that the expressions in the theory are interpreted 'as goals and beliefs'; but they are interpreted as contentful states, of perhaps some other kind, *in the light of* our attributions of goals and beliefs, to the system as a whole.

This is unsurprising, given Dennett's instrumentalism, but he also provides elsewhere (Dennett 1978 107) a persuasive example of a chess-playing computer, which, to the extent that the intentional stance works with it at all, can be said to believe that it should develop its queen early, whereas there is nothing in its program that *says* this; it simply 'emerges' from the actual instructions. Dennett accordingly thinks that

> the only similarity we can be sure of discovering in the *illata* of sub-personal

> cognitive psychology is the intentionality of their labels. They will be characterised as events with content, bearing information, signalling this and ordering that (1981a, 55).

And this is where the crucial dependence on the uppermost-level attributions comes in; this 'content' has to be *attributed* to the postulated events (they have to be *interpreted; cf. also Smith, op. cit.)*, and this can only be done by bearing in mind (however implicitly) the intentional characterisation of the system or organism as a whole:

> In order to give the *illata* these labels, in order to maintain any intentional interpretation of their operation at all, the theorist must keep glancing outside the system, to see what normally produces the configuration he is describing, what effects the system's responses normally have on the environment, and what benefit normally accrues to the whole system from this activity. In other words the cognitive psychologist cannot ignore the fact that it is the realization of an intentional system he is studying on pain of abandoning semantic interpretation and hence psychology (56).

To the extent, then, that SPCP, or cognitive modelling in general, aspires to speak seriously of *representations*, it must depend on the assumption central to the notion of intentionality, that the system being modelled is rational. What it is important to recognise, however, is that this does not entail that its behaviour is always 'correct'. Stich (1982) criticises Dennett's defence of the rationality assumption, on the ground that it shows many psychologists working on reasoning to be wasting their time, since 'all of this work limning the boundaries of human rationality is simply incoherent' due to the fact that 'the presuppositions of intentional explanation . . . put prediction of *lapses* in principle beyond its scope' (54). This criticism is partially misguided because, although there *is*, as will be argued later, a sense in which it's incoherent to look for the boundaries of human *rationality*, and although it's true that intentional explanation fails on errors, what one ought to be saying is that psychologists such as those Stich mentions are working within SPCP, at the *sub-personal* level, where explanation is not (or not only) intentional even though it presupposes an intentional account at a higher level. SPCP, as a previous quotation showed, can, does and should address *errors*; it can and does concern itself with the question of *competence*; but it *cannot* address the issue of rationality.

In the next chapter, we shall follow up these themes by investigating the notion of competence for a theory of this kind. After that, subsequent chapters will raise more general and problematic issues about the relationships between rationality, logic and competence.

4

Competence and Performance

The account of cognitive theorising, just given, would seem to suggest that, in the case of a model of reasoning, we require to postulate a particular account of the processes or algorithms that are responsible for the production of the conclusions offered by subjects to given premises. For simplicity, imagine that we begin by restricting ourselves to the consideration of some well-defined and closely constrained type of task (eg. the solving of syllogisms, or Wason's selection task, or the like), in which subjects are asked to provide responses to stimuli easily seen (at least by the experimenter) as clear and explicit sets of premises. Regarding the premises as 'inputs' and the conclusions as 'outputs', we need to postulate an algorithm which is input-output equivalent to the behaviour of the subjects. As noted, this sort of equivalence is in general far too weak for the aims of cognitive psychology, yet even so it involves a number of serious methodological issues. In the present chapter, I shall aim briefly to expose what I see as one of the main questions encountered in the generation of this input-output characterisation. It will emerge as the nexus of the issues discussed in the last chapter and in the next, and hence will receive considerable further discussion. The present effort, then, is in order to establish sufficient background behind the use of certain terms to prevent their subsequent appearances from being too isolated and uncertain in their connotations.

We have supposed already that the actual behaviour of the subjects may, in some cases, be due not directly to the algorithm, but to its interaction with various comparatively unprincipled interfering factors. There may be behaviour which the algorithm would have produced but for these factors, but in the event did not. A firm line, as on a graph, the algorithm has the subjects' behaviour dotted about it: it must be seen as input-output equivalent only to an *idealisation* of the subjects' behaviour. We can thus expect counter-factual predictions and unpredicted behaviour, and amongst these the question arises as to how we are to select the 'best straight line', the best idealisation. Another way to put this, common in the literature on reasoning (and of course linguistics), is to ask after the nature

of the *'competence'* underlying the subjects' 'performances'. A suggestion often advanced is that people are always essentially *capable* of arriving at a normatively correct conclusion to a reasoning problem, but that for various reasons (poor memory, lack of motivation, distractions, etc.) they commonly do not. Accepting this would entail the postulation of an algorithm which computed a function from input premises to conclusions correct by logical canons, in a fashion compatible (given many assumptions about the sort of system it is operating in) with its failing in the sorts of ways in which people do.

4.1 An anlogy with Grammar

We may wonder, here, why it is supposed that people have perfect competence, how it is decided which normative system best characterises it, and whether they may be right who argue that competence is in fact not perfect. We shall address these matters in the following chapters, but whatever the answers a possible problem, vigorously propounded by Johnson-Laird (1983), is that there is in general a very large (indeed, infinite) set of normatively sanctioned conclusions to any given set of premises, although many of these are trivial (eg. conjunctions and disjunctions of repetitions of the premises, etc.). We naturally do not want to be in the position of predicting that subjects will actually draw all of these conclusions. Johnson-Laird holds that a theory of competence should predict the drawing only of those correct conclusions that people actually would ideally draw, and he undertakes to produce one for syllogistic reasoning (see Part II of this thesis).

His objection appears odd, however, since his analogy is with Chomsky-based psycholinguistics. An extensive quote seems worthwhile:

> The theory of grammar...characterizes the syntactic structure of sentences, and the theory of parsing specifies an algorithm for computing that structure. This approach to psychology is powerful though double-edged, in that a correct theory of competence is an invaluable constraint on theories of performance, but an incorrect theory of competence can seriously mislead researchers. In the case of inference, no one has successfully formulated what exactly the mind computes, and the resulting theoretical gap has been filled by a largely tacit (and accordingly potent) assumption that formal logic constitutes the theory of competence. The fundamental shortcoming of this doctrine is that most inferences in daily life depend on drawing spontaneous conclusions, and reasoners do not draw just any valid conclusion - and sometimes do not draw a valid conclusion at all. It is therefore misleading to assume that what has to be computed is the set of valid deductions, since spontaneous valid deductions are only a subset of this class. (Johnson-Laird 1983 396-7.)

If there is a shortcoming here, it is surely manifest also in the theory of grammar. Most discourses in daily life depend on uttering spontaneous sentences, and speakers do not utter just any grammatical sentence - and sometimes do not utter a grammatical sentence at all. Notoriously, there are 'grammatical' sentences (eg. very long ones) which no-one will ever

produce. The speaker may (in theory) *understand* any grammatical utterance - but then perhaps a subject will, equally, *accept* any normatively correct inference (although in fact, of course, this acceptance, like understanding, is often not immediate).

It is noticeable that the typical method in theoretical linguistics is not the analysis of spontaneous utterances, but rather the elicitation of explicit grammaticality *judgments* on presenting sample sentences to speakers assumed to be competent: Johnson-Laird is frequently scathing about the procedure of offering both premises and conclusions to subjects for validity judgments, and indeed a quite different spread of results is often thus obtained. The point is, though, that the competence theory might be thought to define (in both logic and linguistics) the set within which the predicted utterances ought to fall, if there are no errors, and not the particular responses produced. That this set is unnecessarily large is an interesting suggestion, but surely one which applies similarly to both cases. Johnson-Laird's analogy is a good one, but his own use of it is misguided. What we should have made clear above, therefore, is that people are capable of arriving at *some* normatively correct conclusion, but from the competence theory alone we know not which. Competence theories are acceptors or filters, not generators.

4.2 Validity and Competence Separated

Is there, then, no harm (nor any psychological implications of the sort Johnson-Laird has in mind) in admitting that logic can provide us with a competence theory for reasoning? There appears no special 'theoretical gap' in this particular case, if no competence theory shows 'what exactly the mind computes', if what a competence theory does is show us what *sort* of thing has to be computed; more to the point, perhaps, what sort of thing ought *not* to be.

But perhaps this counter to Johnson-Laird is weaker than it might seem. What he really has in mind is a principled restriction on the actual set of 'correct' conclusions our theory allows. He has a notion of the 'semantic content' of the premises and conclusion, and his suggestion is, in effect, that we should enshrine its preservation, like that of truth, as a basic principle of the competence system (along with the rule that the 'parsimony' of its expression must be increased). He inserts a wedge between the notions of a *valid* conclusion, and a *competent one*.

His grounds for this are ostensibly empirical, derived from observation and analysis of what kinds of conclusions people actually, spontaneously draw. We rely here on evidence from 'competent spontaneous reasoners' and derive our account of competence from this. The rationale and justification for this type of procedure will be discussed in the next chapter, but now we need to contrast it with certain others of Johnson-Laird's remarks. He has a set of criterial goals for any theory of reasoning, one of which reads as follows:

> *The theory must allow that people are capable of making valid inferences, that is, they are potentially rational.* A theory is inadequate if it permits valid inferences to occur only by chance or as a result of processes that do not suffice to establish validity (*op. cit.* 66, original italics).

The significance of this claim is far-reaching, and it will be examined further in subsequent chapters. For the present, though, it discloses a curious two-stage nature in Johnson-Laird's notion of competence. Where people 'do not draw a valid conclusion at all', this cannot be because they are essentially incompetent to do so, but rather (assuming they do not lack motivation, etc.) must evince the effect of some performance limitation; the requirement is introduced *a priori* that competence lies within logic. No evidence is going to be allowed to undermine this requirement, but observational study has its place in restricting the theory to the prediction only of some subset of all valid conclusions. One might have supposed that on this view the latter enterprise was part of the description of *performance*, but Johnson-Laird thinks not. One reason for this, perhaps, is that (mistakenly, as I shall argue later) he regards the existence and nature of mathematics and formal logic as empirical grounds for the quoted criterion, and hence concludes that his entire competence theory is empirically based. But this, again, raises questions. It is clear enough that mathematics and logic require competence for behaviour far removed from anything seen in 'spontaneous' reasoning. A logician might well say 'John is tall: therefore, John is tall or Mary is small', although inferences of this form are certainly not to be anticipated in ordinary discourse - even (usually) that of off-duty logicians. There are therefore apparently *different competences* for these two cases or domains, albeit one is perhaps merely an extension of the other.

This is not a consequence of his theory which Johnson-Laird either mentions or shows evidence of desiring. Nonetheless, I do not wish to deny that it is reasonable: but if it is, where is the ground for insisting that the competence for spontaneous inference cannot be so impoverished as completely to exclude, for some arguments, all valid conclusions? Or, if people must have the competence to draw correct conclusions, say to all syllogisms, why can they not have the *competence* to conclude 'A or B' from 'A'? Do we want to say that their actually doing this would constitute an *error?*

4.3 Validity and Competence Reunited

In order to avoid his wedge being squeezed out here, Johnson-Laird is going to have to assault some traditional intuitions. Separating the dichotomies 'valid/invalid' and 'competent/incompetent', while continuing to equate the latter with 'correct/incorrect', would leave an embarrassing remainder of valid, but incorrect, incompetent conclusions. What Johnson-Laird wants to say, presumably, is that these are, no doubt, logically correct, but they are somehow pragmatically (in a sense captured by his 'semantic content' theory)

incorrect and, in any event and most importantly, they are *not computed.* The problem here remains his insistence that competence is decided in this odd, two-stage way, its logical impeccability being known in advance, while its pragmatic component has to be discovered observationally.

One way out of this is to withdraw the wedge after all, and admit that the sorting out of the pragmatic matters *is* something most appropriately to be placed in the area of performance. The pragmatically incorrect conclusions no doubt are not computed, but perhaps the fact remains that (in suitable circumstances) they might have been, or would have been, which is just what Johnson-Laird wants to say about valid conclusions to reasoning problems for which none such is ever normally offered. On his account, validity is the outside boundary, as it were, of correct performance in any circumstances whatever, and this is just what is generally thought of as competence: there is then only one theory of competence after all.

An alternative is to drive the wedge so far as to create a complete split. In fact, the empirical approach to competence assessment would suggest that the persistent failure to offer any valid conclusion to particular premises can no more be overlooked than the failure to offer certain ones to others. If there are different competence theories for different domains, there seems no obvious reason why they should assume some hierarchical and nested nature, rather than being subject to wholesale revision. In that event, we should have to have a different cognitive model for inference in each domain, with a different central algorithm in each case; and this is a possibility which will be considered in the following chapters. The important point is that (in a domain) the competence theory specifies the behaviour of the ideal subject; the empirical approach simply takes it that the best route to this is the examination of actual behaviour, and admits the possible existence of more than one domain. It accepts also the possibility that competence will be found to have no clear connection with validity, and that the subjects may therefore have to be characterised as 'irrational'.

What I want to do is not to take either of these courses, but rather to leave the wedge more or less where it is. This involves supposing that Johnson-Laird is right in certain respects - in particular, in his view that some aspects of competence are decided on the basis of what is normatively sanctioned, and other aspects on the basis of what behaviour is observed. We do indeed import a large *a priori* component into drawing the boundaries of competence, and this depends importantly on our notions of what a rational agent should be capable of. We are not inclined to wait upon the evidence of behaviour before making assumptions about the abilities of ideal subjects. On the other hand, competence theories *are* intimately connected with the prediction of behaviour, and so I suppose also that it was correct to conclude above that competence theories should then be domain-relative, and

possibly independent of each other. This is more complicated even than it might seem, though, because of the argument to be introduced later in favour of thinking that our choice of competence theory can have a reciprocating rebound effect on ideas of what is normatively right in a given domain. My suggestion is that the instability of the situation leads to an increasing tendency to push the wedge out, but by drawing the notions of validity and competence together rather than simply collapsing the latter into the former. There is as it were an iterative process which modifies the notion of normative correctness (validity) in the light of the emerging notion of competence, and *vice versa.* Naturally, this will need some arguing, and such is part of the burden of subsequent chapters.

Johnson-Laird's difficulty, in particular, now appears as arising from his failure correctly to identify the relationship between competence and normativity, so that he is unable convincingly to recover from the counterintuitiveness of cutting them apart. Although his proposal to involve semantic content as a criterion is intriguing, it remains unclear how it stands in the context of this problem. My conjecture, which Johnson-Laird might be able to accept in part, is that if it became established (in certain domains, eg. some areas of ordinary discourse) that people's behaviour was best described by a theory based on some logic modified by the semantic principle, then there would be pressure to adopt that logic as the ideal of normative correctness in that domain. I depart from Johnson-Laird in conjecturing further that there could be good empirical reasons for including in this normative theory inferences which in other circumstances might be regarded as clearly erroneous; it might not be simply a subset of normally accepted logic. I do not believe that it is possible, at present, further to refine my notion of a 'domain', but I don't see the matter as pressing, since there are many other aspects of the idea that need equally to be worked out.

In summary, we have the following idea. An ideally rational being will, given a set of premises, always advance to a normatively correct conclusion. Assume that humans are, at bottom, ideally rational; to the extent that it captures rationality, normative logic then presents a characterisation of their most competent reasoning behaviour. An algorithm explanatory of that behaviour will therefore be constrained, by that logic, in its input-output characteristics (the corresponding output must be among the logical consequences of a given input). Further assumptions (to be empirically assessed) about the nature of the algorithm and the 'functional architecture' will lead on this basis to an account of the actually observed behaviour. It might be that, to some premises, people *never* produce a correct conclusion (a position that in fact arises with respect to Johnson-Laird's data for performances on one particular syllogism): nonetheless, the algorithm must predict that, in principle, they would have done, had something not interfered with its execution. The internal complexity of the algorithm should, of course, be such as to explain as naturally as possible the errors that are observed to be made, given the postulated nature of the

implementation. It follows that the relativisation of all this to a domain depends on the view that in different domains, different behaviour is rational.

4.4 The Machine Analogy

This could all be expressed by comparing the reasoner to a computer intended to perform inferences, along the lines discussed in the last chapter: the machine is *designed* (programmed) so that it always produces a correct inference; if it does not, there has been some *malfunction*, it has not performed as it is designed to do. There might, of course, be systematic and inductively predictable such malfunctions; the machine might have limited resources, be in an environment which regularly disrupts its operations, and so forth. This does not matter. All that concerns us in our present enterprise is to describe the program, the algorithm, which essentially governs its activity; we are after a *computational* account of its behaviour. The various interferences are not computationally to be described - in retranslation to the human case, they are not *cognitive* processes. Zenon Pylyshyn has put this well:

> Errors and imperfections are not the primary phenomena to be accounted for; rather, it is the competence to deal with the task. The importance of error data is that they provide clues as to how this competence may be realised within certain kinds of resource-limited mechanisms. In other words, errors tell us something about the way the algorithms and the architecture fit together - just as errors in real computers can do the same (though in that case they are called 'bugs' and we try to eliminate them). (Pylyshyn 1978 124.)

The decision as to what parts of the observed behaviour are competent, therefore, is centrally connected with the question as to the basic nature of any cognitive account of it. The central algorithm determines, effectively *is*, the competence theory. Pylyshyn, also, is aware of this:

> The way in which experimental data are interpreted is extremely sensitive to one's implicit normative system. Whether behaviour is to be construed as appropriate to a task or aberrant depends on one's understanding of the task and its goals as well as on the normative system one adopts. (*loc. cit.*)

Thus it is vital to the cognitive theorist to clarify the issues about what the competence theory for a particular domain actually is, and Johnson-Laird is at least correct in asserting that too little attention is typically paid to this question in the psychology of reasoning. It tends to be assumed (tacitly, as he says) that some mainstream notion of classical logic naturally provides what is required. There are two principal doubts about this: Johnson-Laird has already claimed that formal logic will not do as a competence theory; this seems to depend on the domain in question, but what he ought to have said is that in general no standard system such as classical first-order predicate logic is a suitable choice. There will always be *some* formal theory, expressible as a (perhaps highly deviant) logic that

describes the behaviour, but in any event the question remains as to *which* logic is the appropriate choice. In view of the above discussion, we can note also that it looks as though different logics will be found best suited to different domains.

4.5 Competence and Computability

It's clearly possible to think of the algorithm as a proof system. If the competence theory defines the set of 'valid' arguments, then what we seek is a system that is capable of generating (in the ideal case, of course) all and only these: a sound and complete proof system. This is problematic, however. Logicians well know that not all arguments can be said to belong to such a system, as indeed any that go beyond first-order predicate logic do not. Moreover, if we are insisting that there is a *mechanical procedure* (ie, an algorithmic one) for generating these proofs, we are insisting that its basis is a *decidable* logic - but little beyond the monadic predicate calculus fulfils this requirement (though we shall have cause to note later that this includes syllogistic logic).

These facts make it difficult to sustain the notion that a theory of reasoning can be simultaneously algorithmic, and have a competence requirement described by a logic which may not be complete. This is something that Johnson-Laird comes up against when he considers inferences such as (1983, *op. cit*, 140)

> More than half the musicians were classically trained
> More than half the musicians were in rock groups
> *Therefore:*
> Some of the musicians were both classically trained and in rock groups

which cannot be formalised in first-order logic, and apparently not in any complete (let alone decidable) system. He says that the solution to this problem is to be found in effective procedures (ie, algorithms) for translating such arguments into mental models, and then manipulating these to yield up the conclusion (145). It certainly seems obvious enough that the particular example in question can be catered for, and doubtless many more. But if Johnson-Laird wants to claim as his goal the description of a system deriving all valid inferences, it is bound to be impossible in general, since if (say) second-order logic formalises the set of valid arguments of this type, then any effective procedure for deriving such arguments renders that logic decidable, which we know already that it is not. In cases where arguments cannot be formalised by a decidable logic, it is therefore necessary to abandon either the idea that the processes must be algorithmic, or the condition that they (can be shown to) produce all and only valid inferences.

It might be thought that such a result vitiates the substance of the foregoing arguments about competence and normative logic (and the ones that will follow). This would be a mistake, I think, because it is wrong to suppose that the central algorithm in a

cognitive reasoning theory has (provably) to provide a decision procedure for the logic describing the related normative theory. A competence theory, on our account sketched above, is a theory which lays down what inferences a reasoner will draw, in ideal circumstances, from a set of premises. It is based in large part upon the evidence available from experimental and observational studies on reasoners in a particular domain. Such a theory is distinct from a normative theory for the same domain, but is to some extent conditioned by it because the decision as to which of the data should be treated as competent can be only partly determined by considerations of, eg, simplicity in explaining the data. The normative theory is itself reciprocally affected by the competence account, and the two may in the long run converge, but it is clearly wrong simply to identify them. It does emerge, however, that insofar as a competence theory for an essentially algorithmic system is a logical one, it has to be of a decidable logic. I am inclined to regard this as important evidence in favour of the view that one should not conflate competence and normativity, but also as an indicator of a deep problem in the whole area of cognitive theorising about reasoning. For it shows that there is a limit to the extent to which formal theories of competence and normativity can be expected ever to converge, at least for most domains. In particular, it shows that Johnson-Laird can't both have his cake and eat it when he wants to have a theory offering effective procedures for the making of all the valid inferences people can make. This is a point which will receive considerable further discussion in Part II (chapters 12 and 13).

It is worthwhile mentioning here, somewhat parenthetically, a relatively early discussion of some of these issues by Pylyshyn (1973). His concern is primarily with competence theories in linguistics, but he specifically addresses the undecidability problem, and concludes that it is of little psychological importance because competence theories say so little about the details of how tasks are performed. This, while true enough, is no help in dealing with the question of how competence relates to rationality, because that question doesn't depend on the detailed 'psychological reality' of the competence characterisation: its I/O features, so to speak, are enough for the difficulty to arise. We do, however, generally endorse in this chapter Pylyshyn's view of a proper account of competence as a deep theory about the abstract structure of the subject's behaviour, in large measure empirically arrived at as an explanatory construction.

* * *

It might be thought that in identifying the competence theory so closely with the algorithm, I have accepted Johnson-Laird's characterisation of it as a specification of 'what exactly the mind computes'. However, I want to hang on to the idea of competence as being a somewhat abstract notion, not so much a generative one. I suppose this is best put by saying that it specifies what exactly the mind *can* compute. Algorithms in this sort of

theory ought ultimately to be highly complex and sensitive to a large range of circumstantial influences on the reasoning task (any which are cognitively effective, in fact). There may very well in general be more than one computable conclusion to a given inference problem, and no doubt a sufficiently well-elaborated theory might provide an algorithm that predicts which actually will be computed given the particular complex initial conditions of a specific case. We don't want to hold, though, that this is *the* competent conclusion. At the current stage of theoretical development, we want to say that a competence theory is in fact a regulative notion derived jointly from behavioural observations and normative theory, which informs us in our attempts to construct a suitable algorithm. Which, of course, is really just how Johnson-Laird uses the notion.

This allows us to account for the possibility that the algorithm might include some arrangement for taking resource limitations into account. If a machine knows it has a small memory, at can perhaps tell when its usage is becoming critical, and presumably then it will do something sensible such as issue an appropriate message. It will thus avoid falling into reasoning errors of a certain kind, but also its 'competence', as defined by the algorithm, will become itself resource-limited. In this case, I think we still want to speak of its competence as before, recognising that a particular sort of gap has arisen between this and the algorithm that describes its actual behaviour. The case is somewhat analogous to that of a calculator whose 'competence' is described by arithmetic, but whose performance is limited so that after a certain point it produces a warning of imminent overflow. A true *malfunction* might cause errors without warning; the algorithm in the machine specifies what it will do when it's working properly, and although this is consistent with arithmetic only up to a certain point we still want to say that *arithmetic* is what it is doing. That is to say, the theory of arithmetic provides our best interpretive account of its behaviour, and we want to describe the algorithm as being directed at implementing that theory, however imperfectly the available resources allow it to do this.

* * *

In order to pursue the various issues raised in this chapter, it seems sensible to proceed in the direction of an inquiry into the nature of rationality and its relationship to the foundation of normative systems. The general argument here has been in agreement with Ellis (1979), who has the notion of scientific laws being specifications of ideal behaviour, any deviation from which constitutes an 'effect' requiring independent explanation.

> In general, physical laws do not describe the actual behaviour of ordinary physical systems. They say only how these systems would behave if they were ideal. Similarly, we should not expect the laws of rationality to describe the actual thought behaviour of ordinary human beings. Rather, we should expect such laws to describe the thought behaviour of ideally rational beings, and to provide

> only a framework for explaining how ordinary human beings think. Given such a framework, the effects to be explained are deviations from the rational ideal. (Ellis *op. cit.*, 20.)

This 'framework', it has been suggested, can be thought of as a normative theory, but one under strain because ultimately there has to be some close connection between the abstract idea of an ideal being, and the theoretically elucidated nature of a real one. The ideal cannot be too far away; its formulation must take the real into account.

In this context, the reader is entitled to some further explanation (or at least discussion) of what the 'rational ideal' is, how we can decide what it is, and why we suppose that humans even approximate to it. These are the primary topics of the next chapter.

5

Rationality, Logic and Justification

The account of cognitive theories of reasoning, whereby they are characterised as computational in the manner above outlined, draws together several strands of significant philosophical discussion. If the behavioural description of the computation is held to be constrained by an account of the competence of the reasoner, and that competence is identified with some normative logical theory, then we have a natural way of defining 'incorrect', 'erroneous' or 'incompetent' behaviour. But, the question arises, what is the *justification* for characterising these lapses from the competence model as normatively *wrong*, in this way? If there were a simple answer to this, or one relying solely on premises drawn from a different domain of inquiry, it would be none of our concern. However, such is not the case. It appears that as well as *depending* on a normative evaluation of behaviour, our cognitive theory also *contributes* (potentially, at least, in a crucial way) to the grounding of that evaluation.

5.1 Historical Introduction.

The connection can be traced back to Nelson Goodman's account of how normative logical systems are justified:

> Principles of deductive inference are justified by their conformity with accepted deductive practice. Their validity depends upon accordance with the particular deductive inferences we actually make and sanction. If a rule yields inacceptable inferences, we drop it as invalid. Justification of general rules thus derives from judgments rejecting or accepting particular deductive inferences . . . The point is that rules and particular inferences alike are justified by being brought into agreement with each other. *A rule is amended if it yields an inference we are unwilling to accept; an inference is rejected if it violates a rule we are unwilling to amend.* The process of justification is the delicate one of making mutual adjustments between rules and accepted inferences: and in the agreement achieved lies the only justification needed for either. (Goodman 1965 63-4.)

These brief remarks provided the foundations of an orthodoxy. Rawls (1971), in

producing a similar account of the justification of ethical norms, gave Goodman's 'agreement' the felicitous new label 'reflective equilibrium'. Soon after Goodman wrote, however, psychologists began to gather evidence that human performance, on many diverse kinds of task, is not what it might be. People err, often and badly, when compared with normative theories such as logic, probability and decision-making calculi. This is (*prima facie*) odd, if people's practices are supposed to be in reflective equilibrium with those theories.

Some psychologists began to claim that man was not so rational an animal as had been thought; but, on the reflective equilibrium theory, how was this possible? A further complication resided in the idea of Quine and Davidson (among others) that the very basis of ascribing intentional states to others, and evaluating the meaning of their utterances, is a crucial assumption that these others are essentially rational, indeed, logical. Was the psychologists' contention necessarily false?

Richard E. Nisbett (with various colleagues) has appeared as something of a catalyst in a certain corner of this debate. His claims about the demonstrability of human irrationality (eg, Nisbett and Borgida 1975; Nisbett and Ross 1980), as well as those of others, provoked a strong reaction from L. J. Cohen (1981), who argued from reflective equilibrium that human competence was necessarily in accord with justified norms. The storm of protest from commentators reflected the trend behind an attempt by Stich and Nisbett (1980) to replace Goodman's theory with another - appealing, in this case, to 'epistemic authorities'; Paul Thagard (1982) later produced a more subtle effort, appealing to coherence and efficiency, toward a similar end. Most recently, Thagard and Nisbett (1983) tried to combine these arguments with an attack on the 'strong principle of charity' assumed by Quine and Davidson to be characteristic of translation and even intralinguistic understanding.

In what follows, there will be a discussion of the issues of this debate, with an attempt to relate it usefully to a somewhat wider literature, including that of 'naturalised epistemology' in the philosophy of science, and its preoccupation with 'what ought to be believed'. However, no effort will be made at a comprehensive survey here. The central underlying theme in this chapter will be that there is an interplay between normative theorising and the construction of cognitive psychological theories, and that this relationship is important to the development of both. In short, the two must be interdetermining, each depending for its justification (and sometimes also its construction) upon the other.

5.2 The Argument For Necessary Rationality.

There are, as noted, two strands to the argument that people are and must be rational, and these are to some degree independent. We have (i) the argument from reflective equilibrium, whereby it is supposed that the notion of what is correct stems simply from a description of what is typically in fact done; and we have (ii) the Quine-Davidson view that all comprehension of others as persons (not to put it too strongly) depends on taking them to be 'right in most matters', especially matters logical. Let us outline these in turn, and then consider some objections.

(i) Reflective Equilibrium

L. Jonathan Cohen's recent (1981) vigorous attack on presumptuous psychologists is an ideal example of the first, on account of its uncompromising nature. According to Cohen, the actual behaviour of untutored, naive individuals must betray a competence for logically, normatively, correct behaviour (however often this may be faulted in performance), for the simple reason that the construction of the normative theory depends 'at crucial nodes' on consulting that behaviour. The picture here is remarkably similar to that which one finds in Chomskyan linguistic theories. The 'intuitions' (perhaps only implicitly revealed in behaviour) of what one might call native reasoners, are systematised and regularised using essentially the process described by Goodman in the quotation above.

Ordinary human reasoning, Cohen says, sets its own standards. He argues that intuition-free justification for a formal logical system can be had neither from the 'empirical-inductive' tactic of adjoining a logical system to the rest of science and assessing the merits of the whole, nor from any 'metamathematical' argument, which will give us at best a demonstration of the theory's soundness but not its application to everyday reasoning. One has to consult people's intuitions, eg. to see whether 'if...then__' is a material conditional or, if not, what it is. (*Cf.* Anderson and Belnap 1975 - an example of this consultation in action.) And 'unless we assume appropriate intuitions to be correct, we cannot take the normative theory of everyday reasoning that they support to be correct' (319). Of course, these intuitions have to be idealised, generalised over, abstracted from; and the final formal system has to be seen as applying to real arguments only in a rather dissipated way, 'since the actual judgments and reasonings of human beings occur on particular dates and in particular locations, in a particular causal context' (321). However,

> where you accept that a normative theory has to be based ultimately on the data of human intuition, you are committed to the acceptance of human rationality as a matter of fact in that area (*loc. cit.*)

or, perhaps more trenchantly still,

> we cannot attribute inferior rationality to those who are themselves among the

> canonical arbiters of rationality. Nothing can count as a error of reasoning among our fellow adults unless even the author of the error would, under ideal conditions, agree that it is an error (322).

Among the reasons he offers for psychologists' having been misled into ignoring this, Cohen adduces the notion of a 'cognitive illusion'. The terminology is deliberately reminiscent of other, eg. visual, illusions. What one finds here is that there is a wholly characteristic type of error under particular (non-ideal) conditions, but one which is easily seen to be such, even by the subject, if the circumstances are altered (eg, the light changed in the visual presentation, or a suitable content variation inserted into the reasoning problem). The suggestion is that in certain instances - paradigmatically, the 'four-card selection task' of Wason (1966 etc.), and some cases of 'judgment under uncertainty' (eg. Tversky and Kahneman 1973, Kahneman and Tversky 1974) - people's competence is overridden or obstructed 'by factors like the recency or emotional salience of the existing evidential input, by the existence of competing claims for computing time, or by a preference for least effort' (325). Cohen insists that this interpretation is to be favoured over the more frequently encountered view that subjects use radically incorrect heuristics and assumptions in their cognition, and have no general procedure available for doing otherwise. His opinion here is forced by his holding the strong position that

> possession of a competence...entails the possession of a mechanism that must include...a method of generating additional procedures, corresponding to the proof of theorems or derived rules in [the relevant] normative system (325).

Of course, this mechanism is not often much used beyond the simplest cases, because this 'may require skills that are relatively rare'. Such use depends, according to Cohen, on 'intelligence', where 'intelligence is understood not as the competence that everyone has but as the level of those skills that are required to supply the novel input necessary for the discovery of proofs' (326). (If intelligence is the sort of thing measured by intelligence tests, however, there may be evidence against this - *cf.* Wason's famous 'Mensa protocol', Wason and Johnson-Laird 1972.)

A second line of argument offered by Cohen is one which he expounded in more detail in earlier writings (Cohen 1979, 1980). Psychologists (in particular, Kahneman and Tversky, for whose reply see Kahneman and Tversky 1979) are accused of failing correctly to characterise the subjects' task, in normative terms. They have applied, in this case, the wrong theory of probability to their subjects' performances, and have hence held them in error when, Cohen says, they were not. Cohen has a theory (Cohen 1970, 1977) in which he elaborates and systematises a 'Baconian' notion of probability, which is quite distinct in its rules and theorems, from the usual 'Pascalian' sort. Cohen holds that Tversky and Kahneman's (1973, Kahneman and Tversky 1974) data is best interpreted as showing that their subjects are using or conforming to something very like this Baconian probability

system - and that there is nothing irrational or normatively reproachable about their so doing. Kahneman and Tversky contest this, but there is underlying the whole argument an important lesson which will receive further discussion later. One has to be very careful about specifying the normative theory that characterises one's model of 'competence'.

The most apparently unfortunate aspect of Cohen's theory, which will be discussed in some detail later, is the very immunity to disconfirmation which motivates it. Any phenomenon, with sufficient ingenuity, can be redescribed in terms favourable to Cohen; and stipulations to ensure this gradually weaken the explanatory value of the whole. Consider, for instance, the suggestion that an innate competence may depend for its maturation on a felicitous environment, or need a suitable education fully to realise it (322). We shall have presently to ask whether such a dogmatic position is in fact defensible, insofar as it purports to have consequences for the empirical science of psychology.

(ii) The Strong Principle of Charity

The 'strong principle of charity' seems to have its roots in some remarks of Quine's (1960, section 13). These are concerned with the development of his theory about radical translation, the details of which do not presently concern us. Suffice it to say that Quine's project, at the point where these remarks occur, is to say something about how, in principle, we might work out the meaning of the utterances of a native speaking some language with which we are utterly unfamiliar. He explicitly addresses the issue of 'prelogical mentality': might we discover that the native accepted-as-true sentences translatable as having the form 'p and not-p'? Quine thinks not. He supposes that 'semantic criteria' for truth-functions can be stated in terms of assent and dissent (negation turns a sentence to which one will assent into one from which one will dissent, etc.), and then native constructions which fulfill these criteria can, 'subject to sundry humdrum provisos', be translated as 'not', 'and', 'or', etc. Under these criteria, Quine says, the claim about prelogical natives is absurd:

> Wanton translation can make natives sound as queer as one pleases. Better translation imposes our logic upon them, and would beg the question of prelogicality if there were a question to beg (58-9).

Quine goes on to point out that this sort of charitable approach is manifest even where we interpret other English speakers: if someone answers us 'Yes and no', we 'assume that the queried sentence is meant differently in the affirmation and negation'. In general, 'fair translation preserves logical laws', 'assertions startlingly false on the face of them are likely to turn on hidden differences of language', and 'one's interlocutor's silliness, beyond a certain point, is less likely than bad translation - or, in the domestic case, linguistic divergence' (59).

This doctrine has been taken up and elaborated upon by various of Quine's followers, including his pupil, Donald Davidson. The latter has a somewhat complex philosophy of language and psychology in which the assumption of rationality is central. For present purposes I shall introduce this obliquely, via a recent review article specifically about rationality, by another pupil of Quine's (Follesdal 1982). Here, Follesdal notes that, on Davidson's view, the assumption of rationality is essential for the attribution of beliefs, desires and actions (in general, intentional states). The point is that these attributions are inseparable from the interpretation of a speaker's utterances:

> In trying to understand the other person's actions I attribute beliefs and values to him on the assumption that he is rational...In the learning of the semantic aspects of language, belief and meaning are intertwined in such a way that there are not two elements there to be separated:...we assume that a person is rational and then try to find out what beliefs and values we may attribute to him that are compatible with his being rational (Follesdal *op. cit.* 308-9).

Or as Davidson himself puts it,

> the satisfaction of conditions of consistency and rational coherence may be viewed as constitutive of the range of application of such concepts as belief, desire, intention and action (Davidson 1980 237).

> Since charity is not an option, but a condition of having a workable theory, it is meaningless to suggest that we might fall into massive error by endorsing it . . . Charity is forced on us - whether we like it or not, if we want to understand others we must count them right in most matters (Davidson 1973/74 19).

In the writings of more recent philosophers of psychology it has become commonplace to assert that some notion of rationality is implicit in the commonsense attributions of 'folk psychology' (commonplace, but not wholly universal - *cf.* Stich 1982, 1983). This is a cornerstone in the theories of Dennett (eg, 1978, 1981a,b), which are markedly similar to those of Davidson in many respects, and is given an interesting twist by Paul Churchland (1979). Churchland's idea is that folk psychology can be construed as a certain sort of 'measuring system', on analogy with the way in which quantities are measured in, say, classical physics. Mass and acceleration, for instance, are properties (of objects), the relations between which are the relations between numbers; similarly, beliefs and desires are properties (of persons), the relations between which are the relations between propositions. And just as the one set of relations can be handled with the calculus of numbers (ie. arithmetic), so the other can be handled with the calculus of propositions (ie. logic). It's hard to assess just how far the analogy can be pushed, but it is, at least, a clear presupposition of this system that only rational beings can be accounted for within it, since only such are likely to conform to the relevant logical laws. Irrational, illogical beings would present problems for folk psychology at least as great as those presented to classical physics by the

discovery of what happens to the mass/acceleration relationship at velocities near to that of light, and require at least a substantial overhaul and complexification of the entire theory.

It is thus common ground to the collection of views just mentioned, that an assumption of rationality is central to our interpretation of our fellows' utterances, our ability to make any sense even of their non-verbal behaviour, and ultimately our ability to see them as persons at all. The notion of a person is essentially that of a rational being. If Dennett (1982a) is right, we need this assumption even to make sense of ourselves, introspectively. It would seem that there could be scarcely a better ground than the correctness of these arguments, for supposing there to be some basic incoherence in the claims of those psychologists who impugn our rationality.

5.3 Arguments Against Necessary Rationality.

In these circumstances, it is only to be expected that the psychologists referred to, and their philosophical apologists, should mount an attack upon the preceding arguments. I want to consider, in particular, (i) two attacks on Goodman's reflective-equilibrium theory, and then (ii) an attack on the Quine-Davidson position. After that, we shall have the issues in perspective, and be able to hold an informed discussion.

(i) Attacks on Reflective Equilibrium

(a) Stich and Nisbett

Stich and Nisbett (1980) take as their starting point the quotation from Goodman offered at the beginning of this chapter. Their attack hinges on the claim that Goodman makes tacit assumptions about the ways in which people infer, and that these assumptions are empirical and false (189). They cast the argument primarily in terms of induction and inductive inference, largely because that is the kind of literature in which they find the evidence to use against Goodman's alleged assumptions; but they might equally well have spoken in terms of deduction all along. The problem is, anyway, similar in both cases, and the evidence similarly suggestive. Essentially, they claim, one finds that the rules and inferences which are in fact in reflective equilibrium, are often just the wrong ones. Goodman's account, applied to the psychologists' discovered facts, shows all sorts of clearly invalid principles to be fully justified. Stich and Nisbett provide three examples of fallacies which, they hold, are so entrenched in practice that Goodman would have to allow them.

The 'Gambler's Fallacy' is exemplified by the belief that, if a (fair) die is rolled a large number of times without a particular face turning up, then the probability of that face's turning up on the next roll increases the longer this goes on. In fact, probability theory tells us, the face in question will always have a one-sixth probability of appearing;

but people not only fail to realise this in practice, they have even been known reflectively to endorse the fallacy. Stich and Nisbett adduce a telling quotation from Professor Henry Coppee's *Elements of Logic* (1874), in which the fallacy is committed.

The second case is the lack of grasp of the notion of statistical regression to the mean, as exibited by a large range of subjects (Kahneman and Tversky 1973). When asked about their procedure, subjects offer a rule which has nothing to do with regression (Nisbett and Ross 1980), and 'their non-regressive rule is in reflective equilibrium with their actual inductive practice. So for Goodman, both their rule and their individual inferences are justified.' (Stich and Nisbett, *op. cit.* 194.)

Thirdly, erroneous analysis of covariation is considered, as in, eg, Smedslund's (1963) well-known study. People seem typically both to reason in accordance with, and apparently actually to endorse, the obviously defective rule:

> If the presence of A ... is often followed by the presence of B ... then the chance of B occurring is greater when A has occurred than when A has not occurred (Stich and Nisbett, *op. cit.* 195).

This operates, for instance, to the enormous potential benefit of quack doctors; 'often', in the above quotation, may mean 'more often that not', but this is regardless of the relative number of instances observed in which A is present and absent: people also just seem to leave out of account cases where not-A is followed by B, and the other combinations. In inference research generally,

> the picture that emerges...is hardly a flattering one. Subjects frequently and systematically invoke inference patterns ranging from the merely invalid to the bizarre. And, though the evidence is less substantial on this point, there is every reason to think that many of these patterns are in reflective equilibrium (*loc. cit.*).

Stich and Nisbett go on to consider two possible defences that a Goodmanian might offer. Firstly, they suggest (in effect) that perhaps these errors are all due to various kinds of cognitive illusions. In that case, rational subjects can be educated into a better reflective equilibrium, which is more 'stable'. The problem with this is that there is little reason to suppose better reflective equilibria to be the only stable ones (even, they would presumably add, in what Cohen calls 'ideal conditions').

> While it is quite true that subjects can often be gotten to reject invalid rules they had previously accepted, *it is also true that they can be gotten to accept invalid rules they had previously rejected* (196).

Research to prove this has not been conducted in detail, 'for obvious moral reasons'. However, it would surely be churlish to deny Stich and Nisbett their point that 'there is no reason to think that this more stringent reflective process would have a unique outcome for a given subject' (197). What is less clear is whether Goodman would or should be

impressed by this: most likely, he envisaged his 'agreement' as being a rather general one, holding throughout a society or culture, and surely not without many individual exceptions. (Perhaps Goodman's alleged neglect of the 'social component' of justification (189) is more a consequence of his almost cryptic brevity, than anything else.)

The second defence is 'digging in'. Suppose the Goodmanian were to insist that rules in reflective equilibrium for a given subject are thereby justified for that subject. But not for others, and certainly not for *us*. If the idea just attributed to Goodman about cultural norms is correct, then no supporter of his will actually offer this defence; in any event, Stich and Nisbett seem right to reject it as 'bizarre' (198). However, their reason - that it falls short of capturing our intuitive idea of what it means to say an inferential principle is justified - is weak, and calls for clarification. We shall return to it.

Stich and Nisbett follow up their critique with some positive suggestions. The notion of 'epistemic authorities' is adduced: these are just those in society who are acknowledged 'experts' on the subject in question. The suggestion is that deference to such experts is both rational and normatively indicated.

> The man who persists in believing that his theorem is valid, despite the dissent of leading mathematicians, is a fool. The man who acts on his belief that a treatment, disparaged by medical experts, will cure his child's leukemia, is worse than a fool (199).

Well, perhaps. Usually. (As a matter of interest, *Roget* juxtaposes 'disparage' with 'underrate'!) Anyway, might Goodman really have meant that rules, to be justified, should be in reflective equilibrium for the *experts?* Stich and Nisbett think this an unreasonably charitable reading of him; but they think he would still be wrong, even if it were right. The trouble is, they say, that one is not *contradicting oneself* if one holds both that a principle is justified and that the relevant authorities deny it (or the reverse, presumably: Henry Coppee, pilloried above, was no doubt an 'expert' in his own time). And this, surely, cannot be gainsaid. The 'expert reflective equilibrium' is accordingly weakened - Stich and Nisbett present the following analysis of 'Rule *r* is justified':

> Rule *r* accords with the reflective inferential practice of the (person or) group of people I (the speaker) think appropriate (201).

There are *no* restrictions offered on the possible composition of this group. I may unilaterally constitute *anyone* an epistemic authority. Sanity prevails, nay exists, only because 'most people are cognitive conservatives most of the time' (*loc. cit.*).

This view is admittedly 'a bit radical'. Worse, there seems strikingly little, in point of its capturing any intuitions, to distinguish it from the 'digging in' strategy withheld from Goodman's hypothetical apologist. It is scarcely any less bizarre. Moreover, as pointed out by Conee and Feldman (1983), it seems still to entail, counterintuitively, that certain

statements are self-contradictory. Consider

> This rule is justified, but it is not in accordance with anyone's reflective inferential practice (Conee and Feldman, *op. cit.* 330)

- Stich and Nisbett's formula renders this incoherent, but it is surely not so. In fact, their own formula evades *any* of their objections to Goodman's, only if the experimental subjects referred to are 'cognitive conservatives', since otherwise their rules and strategies may indeed be justified. But this consequence is intuitively absurd if anything is, and the assumption (though plausible) is anyway unsupported by any offered evidence. The thesis that the subjects are in any way irrational, or that their practices are in any way unjustified or 'invalid' (earlier taken entirely for granted), thus depends, on Stich and Nisbett's own final account no less than Goodman's, on empirical assumptions, which are neither made fully explicit nor defended.

Despite its ultimate collapse into obscure self-refutation, Stich and Nisbett's paper raises some interesting points, its very failure to illuminate them highlighting their interest. Chief among these is the question: what *is* justification? Where *is* the role of reflection and adjustment, if the statement quoted above from Conee and Feldman makes sense? One response worth considering is that of Paul Thagard (1982): reflective equilibrium has no role; it is 'redundant', 'at best incidental'.

(b) Thagard

Thagard seeks 'a methodology for revising normative (prescriptive) logical principles in the light of descriptive psychological findings' (25). This, he considers against the background of two existing descriptive-to-normative methodologies: historical philosophy of science (HPS), and the theory of 'wide reflective equilibrium' (WRE) in ethics (see Daniels 1979). Neither of these will do, he contends, for the case of logic.

Thagard's characterisation of HPS depicts it as frequently re-running a loop between 'case studies' and 'methodological principles' observed to be upheld in these cases, which process finally eventuates in the production of 'normative models' (27-8). These models are then tested against other case-studies, and so forth. The objection to an HPS-style approach turns on the 'expert' status of the exemplary scientists there considered (eg. Darwin and Newton). Thagard draws an interesting distinction, here, between two sorts of experts: 'those who are expert at performing a task, and those who are expert about explicitly saying how a task should be done' (32). These do not necessarily coincide; we do not have a clear set of experts in logical *practice*, independent of their pronouncements. (This relates interestingly to the reflective/unreflective distinction discussed in chapter 2.) The intended analogy is with scientists: we should look at what they do, not what they say they do, since the latter may be ideologically contaminated (as may the former, no doubt,

but perhaps less or differently so). The problem, for Thagard, is that the practices of logicians (and statisticians, high-level managers and other inferential paragons) are always 'severely infected by philosophical views' (33). There are meta-level debates (such as the classicism/intuitionism dispute or the Pascalian/Baconian problem) which completely condition the inferential behaviour of the disputing factions. Universally valid 'case studies' cannot be found.

The problem is even worse, in a way which Thagard does not mention, when one tries to apply any of this to the question of everyday inference. We do not assume, Thagard supposes, that Tversky and Kahneman's subjects 'know what they are doing'; he thinks that 'instruction in statistics can be expected to change their behaviour in desirable ways' (32). But this is far from clearly the case. Their behaviour would improve insofar as they became statisticians, but they might remain no better off *qua* everyday reasoners. There is no lack of evidence (see Einhorn and Hogarth 1981; Nisbett and Ross 1980; Wason and Johnson-Laird 1972) suggesting that some of the best reasoners, *qua* statisticians or mathematicians, will yet perform poorly *qua* everyday reasoners, if unsuspectingly given a problem to solve in informal circumstances.

This may be a reason for supposing cognitive illusions to be even more pervasive than seems to have been feared. On the other hand, expert everyday reasoners - if there are any - may neither be, nor especially easily become, good formal reasoners. The most effective ways of 'improving' people's everyday reasoning might be by some method radically other than training in the use of any kind of formal rule system. Teachers of second languages are frequently heard to say that it's better to absorb a language through concentrated exposure, rather than attempting to learn its grammar, and this applies also to learning improved writing skills, etc., in one's own language (Lindemann, 1982). The case is no different with logical argument. A good paradigm of cogent, rigorous, yet informal argumentation, is the sort of thing philosophy students are asked to produce. Anyone familiar with such students knows that their production of these arguments is often in no way improved by successful completion of courses in formal logic, and further that the best formalists are commonly a quite distinct set from the best arguers in discourse.

This is related to a point made by Cohen (1981 *op. cit.*; but see also 1979) in which he postulates that there is a 'characteristic indeterminacy' (Cohen 1981: 323, 328) in the experiments of the psychological investigator: when the precise nature of the task has been made sufficiently clear to the subject, for one to be able accurately to characterise what the subject is doing, one has thereby 'trained' the subject to an extent where one is no longer investigating his capacities as an everyday reasoner, but rather as an 'expert' reasoner of some kind. For instance, Cohen maintains that those who commit the Gambler's Fallacy might really be considering the probability, say at the nth toss of a fair coin, of having at

least one tails outcome within any space that consists of *n* outcomes - and this does indeed increase with increasing *n*. Eliminating this possibility entails describing the one in which you are interested, in such detail that error becomes unlikely, or at least has very different implications. (It has become a case of what in chapter 2 was called *reflective* reasoning, whereas previously it was unreflective.) By the time you have made sure you know what the subject is doing, he's probably doing something different; even if he isn't, I would add, he may well be doing the same thing in a very different way. The problem which Thagard has omitted to deal with here, is that one might want to postulate different competence theories for expert and non-expert reasoning (as suggested in the last chapter); and to the extent that one does, the practices of expert reasoners, *qua* experts, are strictly irrelevant to the latter case. This is not, of course, to detract from his point about the difficulty of finding suitable case studies: there are certainly few enough candidates in the literature on everyday reasoning, and on Cohen's view it would be hard to know what to make of them anyway.

* * *

Thagard's view of WRE consists of a loop, similar to the HPS loop, between 'particular moral judgements' and 'ethical principles', the running of which is however strongly conditioned by reference to some set of 'background theories', especially concerning psychological limitations ('ought' implies 'can', so knowledge of this kind may be crucial). The invocation of these background theories is what makes the RE *wide* rather than merely narrow (Goodman's RE is a narrow one). 'Normative principles are outputs from the system only after repeated adjustments of moral judgements and principles in the light of background theories have been made' (Thagard, *op. cit.*, 31). He says that 'unlike HPS, WRE is similar to the psychology/logic problem in that there are no case studies or particular moral judgements with assumed prior validity' (34), and holds that, because of this, the justification of ethical principles is comparatively shaky. Some evidence of 'progress' would be required to make a compelling foundation. However, there are said to be attractive similarities between this model and what is required in the logical case. Background information or theory is essential in the latter also, Thagard holds, and what we therefore require is a wide rather than a narrow equilibrium. We need to know (a) what humans are capable of (since 'ought implies can' holds also in the logical sphere - *cf.* Goldman 1978), and (b) what are the 'goals' of the behaviour. These include achieving true beliefs and avoiding false ones, but also achieving explanations and holistically coherent belief systems. Practical aims are important, and may dictate a cost/benefit trade-off in terms of rigour against psychological simplicity (*cf.* here the remarks of Dennett 1981a about our 'satisficing' rationality, and Thagard's reference to March 1978). But this is not enough, Thagard says: we have, in the case of logic, to take into account a set of background

philosophical theories, and this is something over and above ordinary, ethical WRE (35). It is unclear, at this point, why WRE might not include these, but for our purposes this is an unimportant quibble. What is important is the theory Thagard offers us, and he partially defends this by rebutting Cohen's (1981 *op. cit.*) claim that the reflective equilibrium involved here is and should be only narrow.

Thagard says that Cohen asserts this 'without argument', but this is unfair. It is true, though, that the argument is unclear, and perhaps unconvincing. It depends, in fact, on the remarks, considered above, about the alleged indeterminacy in inferential psychological investigations. The argument is that if we are interested in discovering the competence of untrained people, they are who we have to examine. But any kind of scientific or mathematical reasoning, to which there applied 'normative criteria that were the product of philosophical argument for some appropriately wide reflective equilibrium', could be handled by subjects 'only to the extent that these subjects were not ordinary people but specially trained experts' (Cohen, *op. cit.* 322-3).

> Though a person may well acquire a wide reflective equilibrium with regard to ethical theories that is inconsistent with a previously existing narrow reflective equilibrium, there is no possibility of an analogous inconsistency with regard to deducibility or probability. In the case of deducibility, narrow reflective equilibrium remains the ultimate framework of argument about the merits of other deductive systems . . . (*loc. cit.*)

The motivation for Cohen's distinction here, between ethics and logic, is not obvious. But in any event, he has to be taken as saying that, in the latter case, the emergence of the relevant kind of 'inconsistency' is possible only insofar as the subject has now become a logical expert. And Cohen has claimed already that the systems constructed by such experts are worthless either as competence theories or normative theories, for everyday reasoning, if not shown compatible with the intuitions of ordinary folk. Hence, this kind of inconsistency would show nothing about such folk's irrationality; indeed, Cohen seems committed to saying, it would reflect poorly on the adequacy of the expert's system. Thagard's response to this is to insist that the normative force, central to logical as opposed to linguistic competence theories, is such as to require at least WRE for its foundation. Further, the philosophical issues involved in elucidating the notion of *improvement* in logical practice are such as to go beyond even this. We shall return later to the theme of the analogy between logical and linguistic competence, and in particular the alleged disanalogy in point of normativity.

Thagard offers a model for resolving normative disputes, called FPL ('from psychology to logic'), which is as follows:

(1) We do empirical studies to describe inferential behavior.

(2) We generate sets of logical principles which explain and justify that inferential behavior.

(3) When inferential behavior deviates from logical norms, we consider whether new norms are needed or whether we can just revise inferential behavior to bring it in line with existing norms.

(4) This consideration is based on an attempt to develop a maximally coherent set of beliefs about peoples' actual behaviour, their optimal behaviour given their cognitive limitations and the goals of inferential behaviour, and background philosophical issues.

(5) The logical principles among the maximally coherent set of beliefs are then deemed to be justified.

A caveat is that this way of laying it out conceals the point that there is a loop between 'inferential practice' and 'logical principles', run repeatedly but in the light of the background theories and goals (36-7). As Thagard observes, we now require, with regard to step 4, 'an account of how to evaluate coherence among practice, principles, goals and background theories'. The existence of a set of criteria for doing this, the claim goes, 'renders any discussion of reflective equilibrium redundant' (*loc. cit).* Thagard proposes three main criteria: robustness ('a system is robust if its normative principles account for inferential practice in a wide range of situations'), accommodation ('we can accomodate [deviant] behaviour by using background psychological theory to explain why in some cases people deviate from the system's logical norms') and efficacy ('the extent to which the principles and practices of a system lead to satisfaction of the relevant inferential goals').

An 'inferential system', Thagard says, is a matrix of four elements: normative principles, descriptions of inferential practice, inferential goals, background psychological and philosophical theories - schematically,

$$S = \langle NP, IP, G, T \rangle.$$

Thagard then maintains that

> In a given domain, we can assume that T and G will be common to competing systems, and this gives us some hope of reaching an objective conclusion that one system is more coherent than the other. In particular, comparison of the *efficacy* of the two systems may enable us to make choice of systems more than a matter of purely internal coherence. Choice will obviously be highly complex and non-algorithmic, but nevertheless may be determinate and objective (38).

'In a given domain...': here we need to know (but are not told) what is meant by a 'domain', since otherwise it is unclear why the stated assumption is unassailable. For instance, it is rather far from evident that T and G are or should be the same for systems characterising the inferences of everyday reasoners, on the one hand, and those of various

kinds of experts, on the other. Are these, then, different domains? If so, what are the consequences for those who would impugn the rationality (in the sense of normative-system adherence) of the everyday reasoner? In particular, does a theory have to be 'robust' only relative to a domain? Thagard obviously supposes, when FPL is originally defined, that T and G are universal; that there just are certain essential goals, and perhaps also psychological theories, underlying all inference and inferential behaviour: but the first of these ideas is unargued, and the second flies in the face of several psychological observations, as noted above and in chapter 2. One suspects that in fact the notion of a domain is intended to be nothing more significant than the distinction between deduction and induction or decision theory: even here, though, the characterisation and delimitation of domains may be tendentious - Thagard does not go into it. How many domains are there? How do we decide?

Given his account, Thagard argues, the notion of reflective equilibrium has no work left to do. Stich and Nisbett (*op. cit.*) are cited as evidence that 'resoundingly non-efficacious' inferential systems may be sanctioned by it, including such errors as the Gambler's Fallacy. But is this particular fallacy (at least) really 'non-efficacious'? (*Cf.* Cohen 1981 *op. cit.* 328). The answer is less than apparent because of the implicit feature of Thagard's model, that efficacy is relative to G, the inferential goals of the reasoner. These goals are at best vaguely specified, and it seems that they may hold only over a specific domain. Hence, it may be that domains exist in which the goals of the reasoners are such as to render the gamblers' fallacy (for instance) wholly efficacious and thus justified. It does not help to say that

> the justification of a set of normative principles is based, not on the reflective equilibrium of any individual or group, but on the place of the principles in a defensible inferential system. Defense is based on arguments that the system is coherent according to the criteria discussed above (39).

It does not help because the smuggling in, in a crucial phrase, of the notion of a domain, leaves open the possibility that Thagard's model is also constrained to relativise its characterisation of coherence to the practices of an individual or group, with the unwelcome result that, lacking arguments for the near-universality of T and G, its promise of greater 'objectivity' is somewhat diluted. The analogue to the question '*whose* reflective equilibrium?' is the question '*whose* goals, and the theories of the psychology of whose reasoning?'. No argument backs Thagard's assertion that 'education in sophisticated inferential techniques can be expected to provide the individual with a much more efficacious system' (*loc. cit.*), and such argument would have to show that these techniques *would* be more efficacious, given (a) the individual's goals, and (b) the nature of his inferential psychology, particularly the constraints operating in the circumstances of his typical inferential performances (since Thagard has said:

> Efficacy also should take into account what background psychological theories tell us about how easily principles will be applicable given human information processing mechanisms (38)

- and we are surely at liberty to suppose this will not be constant over different domains and circumstances). Thagard falls foul of the same problem when he follows Stich and Nisbett and urges us to take the advice of experts, saying that they are more likely to have highly coherent inference systems (40): this depends on his claim that experts are more familiar with the elements in the matrix than are ordinary people, but in fact the background theories are as yet in an infant state and we have no persuasive characterisation of most everyday inferential goals. What is most crucially lacking, in fact, is some account of what inferential goals there are or might be, and indeed ought to be, which is something we shall presently be investigating.

These, then, are the attacks on reflective equilibrium. Let us turn now to the matter of the Quine-Davidson argument for the necessary rationality of all people.

(ii) An Attack on Charity

Thagard and Nisbett (1983) present an attack, most generally, on all theories which entail a strong 'principle of charity', that is, theories which insist that without overwhelming evidence to the contrary we should not suppose anyone to be less than wholly rational. The position of those influenced by Quine and Davidson is an extreme example of such a theory, seeming to indicate that we *cannot* suppose someone to be less than wholly rational, at least while continuing to make any sense of him at all. The strategy of this attack is to show that a moderate principle of charity is acceptable, which however leaves room for empirically justified judgments of irrationality.

To begin, rational behaviour is defined as 'what people *should* do given an optimal set of inferential rules' (Thagard and Nisbett 1983 251). It needs to be noted that these rules are not immutable, but subject to change and revision: however, to be irrational is to violate 'the principles that we currently, objectively, hold'. Given this, the most general principle of charity is: Avoid interpreting people as violating normative standards. But this is too vague, so Thagard and Nisbett propose the following five different levels of stringency in such principles:

(1) Do not assume a priori that people are irrational.

(2) Do not give any special prior favour to the interpretation that people are irrational.

(3) Do not judge people to be irrational unless you have an empirically justified account of what they are doing when they violate normative standards.

(4) Interpret people as irrational only given overwhelming evidence.

(5) Never interpret people as irrational.

The proposal is that level (3) is the correct level for principles of charity as canons of social science methodology.

In criticism of Quine, Thagard and Nisbett quote some of those passages from *Word and Object* (Quine 1960) which were offered in section 5.2 above, and suggest that these indicate a level (5) principle of charity for translation. Suppose though, they ask, that a native were to assent and dissent in most cases, in such a fashion as to establish plausible translations of logical words such as 'not' and 'and', but then, in isolated cases, apparently assented to blatant contradictions? Would we not have either to abandon our translation, or to admit an inconsistency here? Quine, they say, urges the first option, but surely the second makes more sense (253). At this point, one is tempted to go along with Thagard and Nisbett. It seems clearly absurd to abandon one's hard-won translation in such a case; and indeed a sensible procedure is, one would have thought, to keep the translation only unless and until the frequency of deviations becomes such that an alteration will lead to a simpler system overall. Quine is sensitive to this, however, and as we noted above really suggests that where we have (on our established translation) a *prima facie* contradiction, we need only alter the translation (*qua* theory) if no special story can be told about this particular isolated case. But, of course, usually (as in the 'Yes and no' example) a special story *can* be told, and telling it is a manifestation of charity. Thagard and Nisbett seem on stronger ground, though, in citing Kekes' (1976) ideas about the asymmetric notion of identity apparent in the religion of the Nuer, as described by Evans-Pritchard. It is probably true that

> [u]nderstanding illogical beliefs requires a much larger battery of hermeneutic techniques than Quine's behavioristic reliance on criteria of assent and dissent (*loc. cit.*)

but we must beware of excessive glibness in talk of 'illogical beliefs', and be alert to the consequences of talking of them at all. Thagard and Nisbett's idea, anyway, is that apparently incoherent practices or concepts can have a 'social role', and that 'because language can have functions other than communication of truths, translation cannot always be charitable' (254). One cannot deny that in many ritual practices, as well as much perfectly ordinary poetry, contradictions may abound. But this is unimportant, it seems to me, since the laws of logic, the breaking of which is here supposed to constitute irrationality, *lose their normative force* (if not their very application) in such circumstances. Logic is concerned primarily with the communication of truths (or at least the treatment of propositions stated as such), and is elsewhere of doubtful relevance. It is at the lowest estimate tendentious to suggest that it is *irrational* to break logical laws in non-literal language of the sorts in question. It is not, therefore, an argument against charity to point these out.

Still less is it an argument against Quine since, as I read him, he is saying nothing about any language uses that are at all metaphorical. Hence, when Thagard and Nisbett say

> The argument against level (4) and (5) principles of charity is that we can gain such a thorough knowledge of a subject's language through study of those other aspects [of his thought - eg. at other levels of generality] that, when faced with an apparently contradictory utterance, we should construe it as contradictory rather than revise our well-established translations

the response is that perhaps we should construe it as *figurative*, or else as a symptom of confusion or misunderstanding, in which case the actual existence of a contradiction is not indicative of irrationality. Quine may be regarded as assuming that the subject's assenting etc. can be known to be free of such problems while the translation is being produced: indeed, he says that in cases, eg. of conjunction with long components, 'the subject may get mixed up' (Quine *op. cit.*, 58), and so limits his account to short components. Presumably, Quine's point is that translation has to start in cases where assent and dissent can be regarded as genuine and sincere, cases where the subject is being related explicitly with his environment and statements about the latter are being made and evaluated for truth. If rationality has to be assumed here, it can hardly be dropped with impunity later on.

Thagard and Nisbett follow up these anthropological musings with the problem of Hegel. Hegel apparently challenges the principle of contradiction, but to understand this 'we need a full account . . . of his complex notions of dialectical negation and contradiction' (255). The suggestion is that despite this he can only be understood and translated as violating that principle, which 'is not to be 'uncharitable', but merely to take him seriously as a complex and iconoclastic thinker'. Perhaps so, but it is not to regard him as irrational either. A strain is developing between Thagard and Nisbett's argument against principles of charity concerning the breaking of logical rules, and against those concerning irrationality. Maybe, *contra hypothesi*, irrationality is not thus to be defined. In any event, this quotation from Quine seems to provide a response to the argument about Hegel:

> Consider the familiar remark that even the most audacious system-builder is bound by the law of contradiction. How is he really bound? If he were to accept contradiction, he would so readjust his logical laws as to insure distinctions of some sort; for the classical laws yield all sentences as consequences of any contradiction. But then we would proceed to reconstrue his heroically novel logic as a non-contradictory logic, perhaps even as familiar logic, in perverse notation (Quine *op. cit.* 59).

Hegel's 'complex notions of . . . contradiction' are surely such as to insure distinctions of *some* sort: and it therefore seems likely that Quine's way out will apply. If it doesn't, then perhaps Hegel *is* irrational, and his thought complex to the point of insanity. (The other feature to notice about this quotation is the implication that Quine is prepared to accept that the system builder be bound by an unfamiliar, although necessarily 'non-

contradictory', logic. This is a point that will loom large later.)

The strain referred to in the last paragraph becomes severe at the end of Thagard and Nisbett's section on translation. They offer the following conclusions:

> Behind the translational principle of charity is the assumption that the primary function of language is usually communication . . . But in many contexts, from Zen to cultural rituals, language can be used for social or other ends. Maintenance of principles of logic may be irrelevant to such ends.
>
> Hence translation of the utterances of people of other cultures and philosophies may well *require* that we understand them as violating our principles of logic, of rationality (Thagard and Nisbett, *loc. cit.*).

The assumption at the beginning, here, seems to be that language can be used sensibly and rationally for other ends, and logic is *irrelevant* to this. Translation of the utterances in question depends, not on finding that people are *violating* logical principles, but rather on finding that logical principles simply do not apply to their utterances. They are not *stating facts;* we are not in the position of having to ascribe to them *inconsistent beliefs;* they in fact are not appearing to be in any way *irrational.* So far from arguing that people need not be seen as rational, Thagard and Nisbett have really argued for the conclusion that rationality is not to be determined by 'maintenance of principles of logic', at least in circumstances where the expressing of truths is not in question. This may be unQuinean (especially without that final caveat), but it is not their advertised aim.

* * *

Attention is next turned to the cognitive psychologists' information-processing account of the derivation of new beliefs from old, which is nearer our central concerns. In a rationally organised system, one supposes, this proceeds in accord with logical laws connecting the contents of the beliefs. Is logic 'hard-wired' into us? Must it be? This, they say, is particularly implausible in the case of inductive principles - which is not surprising given that 'statisticians have grave disputes about the foundations of their work' (Thagard 1982 *op. cit.* 33), and thus that we don't really know what these principles are. It is also no news, despite the continuing effort apparently directed at proving it, that 'people make inferential errors', and that they 'are not fully rational in that their inferences fall short of the best available normative standards' (Thagard and Nisbett *op. cit.* 257). *Performance* has always been known to be imperfect by these standards, which is the commonsense basis of much advertising practice, and the like, but surely mere *error* has to be distinguished from *irrationality,* of which it may or may not be a symptom. These facts, though, are now employed in an attack on Dennett's claim that any intentional system must be rational (which Dennett now insists is *not* the claim that such systems 'must be supposed to follow the rules of logic' - Dennett 1982a). There is a substantial problem here,

it has to be admitted, but treatment of it will be deferred until after the current exposition, since some further literature will be dragged in. Let us now merely note that Thagard and Nisbett are probably right to assert that

> There is no reason . . . why it should not be possible to determine empirically that a system is regularly using some inferential principle or heuristic that departs from standard logical principles, then to use the operation of this heuristic as part of an explanation of the system's behavior (*loc. cit.*)

- the question is to what extent this falsifies Dennett's claims, and the answer to be offered will be that even to formulate this sort of account itself crucially depends on a rationality assumption (much as the attribution of inconsistency in discourse depends on a translation based on charitable assumptions).

Dennett is joined by Davidson in Thagard and Nisbett's target area, when the point is noted that both these writers urge not only that inferences are rational, but that in general beliefs are true. Davidson says, we must count others 'right in most matters', which is characterised as 'an astoundingly strong kind of charitable principle' (Thagard and Nisbett 260). Now, of course, we all know that everyone has a large number of false beliefs. This is not at issue; rather, the question must be when, and in what circumstances, we are entitled to ascribe these. What Davidson and Dennett are both suggesting is that the ascription of each false belief must be embedded in a web of true beliefs, or at least beliefs shared with the ascriber, or else the system would be too incoherent to avoid collapse. The context of Davidson's remark (elided from the quotation in section 5.2) is:

> Until we have successfully established a systematic correlation of sentences held true with sentences held true, there are no mistakes to make. Charity is forced on us . . . we must count [others] right in most matters (Davidson 1973/4 19)

and his point is that we must seek to 'maximise agreement' - the process of interpretation or translation depends on these assumptions - and only then will meaningful disagreement be possible. But Davidson also says we must do this

> subject to considerations of simplicity, hunches about the effects of social conditioning, and of course our common sense, or scientific, knowledge of explicable error (*loc. cit.*)

which seems to be precisely what Thagard and Nisbett are calling for!

* * *

In their conclusion, Thagard and Nisbett reiterate their central claim, which is that 'whether people's behavior diverges from normative standards is an empirical question'. They seem satisfactorily to have shown that people's behaviour certainly diverges from *some* normative standards (although, as Cohen suggests, perhaps sometimes irrelevant ones), but then we knew this anyway. The important question is what this shows about

people's rationality. One is inclined to suppose that (a) people may (and often do) diverge from normative principles for mundane reasons to do with confusion and forgetfulness - ie. these are errors of performance rather than competence - and that (b) since logical principles are revisable (as Thagard and Nisbett insist), people may sometimes be diverging from them quite rationally. Surely it is not *irrational* to forget, or to act on principles into which presently accepted canons are destined to be revised. Thagard and Nisbett do not seem to have made it clear that 'strong principles of charity are indeed hindrances to understanding human behavior' (264). Nor are they 'socially undesirable': it is just false that

> the assumption that people *must* be fundamentally rational blocks the possibility of systematic education to eliminate the gap between behavior and normative standards

- indeed, it is not evident that *systematic* education should succeed if people were not rational. It is certain, at least, that the 'gap' in question might both exist and be closable on any commonplace account in which competent reasoners err because of (in educated practice controllable) performance factors.

Thagard and Nisbett's parting shot is plausible only because of their erroneous conflation of rationality with adherence to a particular set of normative principles: they suggest that strong principles of charity will hinder the development of new normative principles. This cannot be, if these new principles are to be decided in the light of what is rational, and if charity is about rationality. The worthwhile point to be taken here is that this danger will arise, if rationality is shortsightedly equated with current logic: and we are forewarned accordingly.

This concludes the advertised survey of attacks on the two ways of presenting the notion of man's essential rationality. None of these attacks was seen to be wholly convincing, and in the role of advocate for Quine and Davidson, we might (almost) expect to see concerted opposition collapse. This would be premature, however: perhaps something can be salvaged from the debris which will illuminate the central problems here. What seems to be indicated is a deeper analysis of the positions assailed, in the light of the pointers provided by the critics.

5.4 Summary

In this rather long chapter, we have begun to look into the question of the relationship between logic and rationality in psychology. Ultimately, the aim is to uncover the way in which, as I have claimed, logic and psychology come to be in a position of mutual determination and justification. This aim will be further pursued in the following chapter, but so far we have examined some of the issues raised in previous discussions in the

literature.

We pointed to Goodman's introduction of reflective equilibrium, and followed this up with descriptions of the work of Cohen and some of his critics. We considered also the view, inspired by Quine, that the assumption of rationality is a central pillar of the network of concepts often referred to as 'folk psychology', especially the language of intentional states and their attribution, including the problem of radical translation. Some arguments were then investigated, that purport to show the inappropriateness of these ideas. Stich and Nisbett's critique of reflective equilibrium was seen to be defective, but still suggestive that something is wrong with a simplistic approach to it; we therefore investigated Thagard's attempt to replace it with a more complex system. The view was taken that this develops a useful framework, but one marred by an over-eager acceptance of the idea that ordinary reasoning should be assessed by the standards of 'expert' reasoning, that stems from a failure to follow up its own pointers to the importance of the domain and the set of inferential goals, with respect to which the reasoner is operating.

It was claimed that Thagard and Nisbett's efforts to reduce the significance of the Quine-Davidson view on 'charity' fail, on account of confusions between error and irrationality, between departing from logical principles and using language to which they are irrelevant, and in general between current logical principles and rationality. The arguments seemed not to appreciate the weight of Dennett's and Davidson's claim that attribution of *false* beliefs depends on a background of a shared system of beliefs held *true*. The task of the next chapter is accordingly to explore these ideas further, with particular reference to the role of rationality in relation to the idea of competence.

6

Rationality Reconsidered

To recapitulate: a crucial question before us is, what is the relationship between characterising a piece of behaviour as 'incompetent' according to some cognitive psychological theory, and characterising it normatively as 'wrong'? We have seen that there is generally supposed to be a relationship between competence theories for reasoning, and 'logic': more often than not, it is taken *a priori* to be the case that this relationship is identity. It is also often regarded as uncontroversial that some mainstream notion of logic - predicate calculus, or (as frequently in the case of syllogistic reasoning) some more or less Aristotelian theory - encapsulates the canons of 'correct' inference, departures from which are *ipso facto* 'wrong'. However, the concluding criticism, just offered, of Thagard and Nisbett, indicates that this is too hasty a way with the problem. Contemporary theories in which logic is seen as subject to reform and progress, rather than as a collection of eternal truths, clearly point to some higher court of appeal on the matter of correctness. This, I am suggesting, is the court of rationality: the correct is the maximally rational, and the justification of a logical system depends on maintaining that its recommendations are everywhere rational. Here, we suppose that on accounts such as Goodman's, or Thagard's, strategies like the search for reflective equilibrium, or the increasing of coherence and efficacy, are just suggestions as to how best the rationality of a system is to be maximised.

The response to the opening question then is that lapses from theoretically 'competent' behaviour are 'wrong' if and only if they are seen also to be lapses from rationality. The traditional type of account is therefore right just insofar as the psychological theory at issue includes an account of competence which is identical to (or at least constrained by) a logical system that successfully captures rational inference. This way of putting the matter still has its disadvantages. It creates the appearance that there in fact does exist a body of eternal truths, this time of rationality rather than logic, to the capturing of which all logical theories aspire; and it is not clear that this gains us anything over what it replaces. The appearance, though, can be dispelled by reflecting on what the notion of

'rational inference' is, if rationality is thereby revealed not to be something eternal or divorced from empirical considerations.

Clearly, at any particular time, the currently accepted canons of normative logic (or decision theory, etc.) are taken to be the most expressive of rationality. The suggestion to be amplified here, however, is that the attempt to use these as a guide in the construction of a cognitive reasoning theory - most particularly, to use them as an *a priori* characterisation of the appropriate competence - may lead to their re-evaluation in this respect. It may be that the characterisation of rationality changes, in the course of this theory construction, such that changes are forced in the normative theory of inference. We shall see presently some more of the implications of this.

6.1 What is Rationality?

In the present chapter, I want to consider what the required notion of rationality might be. A possibility which arises is that the authors discussed in the last two chapters are using more than one such notion, and hence that some of the disputes are somewhat empty. In any event, we need to try to find some notion appropriate to our general discussion in the context of cognitive theories of reasoning, which will leave us in a position to evaluate the impact of the above discussions on our project.

Follesdal (1982 *op. cit.*) notes that Elster (1982) distinguishes more than twenty senses of 'rationality', many of which, however, are neither in conflict nor of any particular relevance to the sorts of issues we are addressing here. It seems sufficient to consider, as Follesdal does, a smallish subset of these. Follesdal settles for four:

(1) Rationality as logical consistency

(2) Rationality as well-foundedness of beliefs

(3) Rationality as well-foundedness of values

(4) Rationality of action.

Follesdal notes that (2) is much stronger than (1), since 'well-foundedness' presumably involves a large number of empirical factors, and that (1) itself is open to different strengths of interpretation, depending on whether it is taken to mean that one has no contradictory beliefs, or that one's beliefs entail no contradiction. The importance of (3) relates to its potential in evaluating the rationality of normative prescriptions, whether in ethics (which is what one initially takes 'values' to imply) or science and logic - this type of rationality, as Follesdal says and as we have already seen, is generally and most promisingly discussed in terms of reflective equilibrium. Follesdal holds that (4) is best thought of in terms of decision theory (by which, however, he seems to mean something more akin to Dennett's 'intentional systems theory'), bearing in mind the epistemic and cognitive

limitations of the particular subject:

> Rationality of action is normally a question of how to make the best use of one's resources, one's information-seeking capabilities, and one's ability to create good alternatives, and not a question of choosing from within a vast set of alternatives that lie there ready for one's inspection (Follesdal 1982 307).

He also makes the point, with regard to (3), that when one is asked to discuss the rationality of an individual, only rarely does one attend to his values, rather than simply his actions and the beliefs from which they stem - perhaps because of the tendency to regard ultimate values as 'beyond the realm of rational justification'. One concentrates only on those aspects of intentional states which are related to their logical form.

Follesdal takes this neglect of the content of states to be a mistake, but it seems nonetheless to be in a sense at the root of the notion of rationality inherent in the Quine-Davidson approach. It will be recalled that mention was made above of Churchland's (1979) way of stating this in terms of intentionally-describable objects or systems (ie, persons) conforming their intentional states to the rules of a logical calculus. Here, no attention is paid to the origins of the system's beliefs and goals except insofar as they are produced from other beliefs and goals. To the extent that a rationality assumption is indeed involved, it obtains just so long as these states are appropriately related, no matter how bizarre they may be in themselves. One might say that the system, and its behaviour, is rational on this occasion or in this respect, *given that goal* (and those beliefs).

Dennett, in 'Beyond Belief' (1982c), endorses Churchland's general approach, but is always somewhat evasive about the relationship between logic and rationality. Stich (1982) takes Dennett to task over some alleged consequences of his rationality assumption (which controversy we shall discuss in due course), and Dennett (1982a) replies by admitting that he has never been very explicit about what he takes rationality to be. He says the concept is 'slippery' and 'systematically pre-theoretical', and remarks,

> I want to use 'rational' as a general-purpose term of cognitive approval - which requires maintaining only conditional and revisable allegiances between rationality, so considered, and the proposed (or even universally acclaimed) methods of getting ahead, cognitively, in the world. I take this usage of the term to be quite standard, and I take *appeals* to rationality by proponents of cognitive disciplines or practices to require this understanding of the notion. . . . [Cf. theories about compartmentalisation of memory:] The claim is that it is rational to be inconsistent sometimes, not the pseudo-paradoxical claim that it is rational sometimes to be irrational. . . . One may, then, decline to *identify* rationality with the features of any formal system . . . (Dennett 1982a 76-7).

This suggests that Dennett's 'principle of charity' does not even extend to insistence on the rule of noncontradiction. It all depends on the reasons for a contradiction's occurrence.

However, it has to be said that others of Dennett's observations on the subject are not clearly in keeping with the above. One of his best known discussions, 'Intentional Systems' (in Dennett 1978), conveys a strong impression that we can assume an intentional system to have its beliefs, goals and actions interrelated very much as Churchland suggests, with the actions (or intentions to act) being derived by some form of implicit practical syllogism from the other states. The idea seems to be that this follows from the system's having evolved successfully (in the case of a natural system) or from its having an optimal design. The implication is that the possibility of adopting what Dennett calls the 'intentional stance' (or, as in 1981b, the 'intentional strategy') towards a system depends on its being organised this way, and that any optimally efficient system in fact will be organised this way. Follesdal also emphasises the point: there would be no way to work out a person's beliefs and goals from the evidence of his actions (*the only possible evidence*) were one not to presuppose a rational organisation. Rationality is constitutive of the notions of beliefs, values and actions.

Perhaps, though, the idea that this kind of rationality is strictly related to logic, is on the wrong lines. Follesdal asks: 'How much rationality do we have to require in order to talk meaningfully about desires and other 'intentional' notions?' (312) - and his answer is that we need enough to be able to indulge in 'reason explanations' of someone's actions, rather than mere 'causal explanations'. It's not clear, however, that this is more than a tautology: 'reason explanations', surely, just *are* explanations in terms of beliefs, desires and other intentional notions. (I leave aside, here, Davidson's, 1963, view that reasons *are* causes.) In any event, Follesdal agrees that one does not often find perfect rationality in sense (1) above, especially interpreted strongly. Rationality in sense (4), on the other hand, seems common, provided we take care not to make unreasonable assumptions about what the agent has taken into consideration.

Follesdal is again slipping here into talking of the rationality also of beliefs in themselves, rather than simply that of their interactions:

> causal, irrational factors come into the picture at a number of places, where our general theory of man makes us expect them to come in, as when a person gets his beliefs and values formed through propaganda, advertising or group pressure, or when he acts under the influence of hypnosis, drugs, drives with which he is not yet familiar, etc. (313).

It is important (if we can relate this to our earlier discussion) to distinguish the features of those processes which are to be regarded as competent - 'due to the (supposed) algorithm' - and those not. To what extent does the algorithm purport to cover the *acquisition* of beliefs? And how far is its operation prey to various interfering factors, such as the relative emotional salience of pieces of information? One might advance the view that Follesdal's 'causal' factors, are just those which give rise to behaviour 'incompetent' by

the standards and operations of the cognitive theory; the domain of 'reason explanation' is coextensive with that of algorithm-explanation. This returns us once more to the position of asserting the essentially perfect rational competence of humans, ascribing faults to 'causal' performance factors. Consider this:

> In trying to understand a person and his actions, we must weigh against one another [assumptions about his beliefs and attitudes] that we make on the basis of observation of his actions and what goes on at his sensory surfaces . . . Causal factors that are not reasons should in part be regarded as influencing the agent's consideration of alternatives, in part as affecting the alternatives themselves, and in part as dispositions that are to be weighed together with the person's beliefs and desires . . . Physiological urges, hypnotic instructions, and sub-conscious beliefs all seem to be amenable to this kind of treatment (*loc. cit.*)

Here again we have the picture of trying to interpret the agent as wholly rational, except where the interpolation of causal factors seems essential.

It might be worth noting that Follesdal apparently does not hold 'reason explanation' to cover a domain coextensive with that of explanation in terms of contentful states (unlike Pylyshyn - *cf.* ch. 3, above - and probably Churchland and Dennett, and certainly Davidson 1982). Hypnotic instruction and subconscious beliefs (and desires) are evidently among the latter, and hence, perhaps, would be better regarded more as among the range of reasons (if bad ones) rather than that of causal factors. The ascription of subconscious beliefs, that is to say, is part of a rationality-assuming reason-explanation. This, it seems to me, extends even to an example Follesdal uses in arguing *against* trying to see people as maximally rational (310, 314*f*). The example (attributed to Suppes) is that of a pubescent boy who explains his keenness frequently to come up and ask questions, after class, of his attractive female teacher, purely in terms of his desire to learn - and does this in complete sincerity. Clearly we are likely, here, to suppose that the boy's explanation is wrong, and that his real, if subconscious, motivation is much less intellectual. But is this to denigrate the boy's *rationality?* Follesdal says 'we should not easily set aside the agent's reason explanation in favour of a causal explanation' (314); but this may not be what we propose to do. We might say, instead, that the agent here is not (fully) cognisant of his own *reasons;* we might say that he acts as he does because of beliefs and (more particularly, in this case) desires which, unfortunately no doubt, he does not himself recognise. This pattern of explanation would *not* be causal. Causal factors, maybe, would account for his failure correctly to assess his own mental states - and here we have an interesting problem for the psychology of what is known as 'cognitive dissonance' - but this does not seem to provide (what Follesdal claims) an argument that we should not ascribe beliefs and desires so as to make the agent come out as rational as possible.

Follesdal offers the principle:

> When ascribing beliefs, desires and other propositional attitudes to a person on the basis of observation of what he does and says, do not try to maximise his rationality or his agreement with yourself, but use all your knowledge about how beliefs and attitudes are formed under the influence of causal factors, reflection, and so forth, and in particular your knowledge about his past experience and his personality traits, such as credulity, alertness, reflectiveness etc. Ascribe to him the beliefs and attitudes you should expect him to have on the basis of this whole theory of man in general and him in particular (316)

- which seems laudable enough (and closely resembles Thagard and Nisbett's suggestions) but which taken with the paper as a whole seems to say: temper your ascriptions with the knowledge that the individual's capacity for perfectly rational actions is tainted by a mortal proneness to causal performance limitations. As observed before, we all know in advance of any high-level discussion, that people's actual behaviour is never perfectly rational, and very often is a long, long way from that ideal - but here we do not have something that Quine and Davidson (or Dennett and Churchland) have failed to allow for.

Given Dennett's quoted proposal to use 'rational' as a term of 'cognitive approval', one might be tempted to argue as follows. Follesdal has essentially given his seal of approval to those kinds of causally (rather than reason-) based infelicities in behaviour that we do not typically avoid, and that we can reliably predict from general theories of man. He has, perhaps, decided to regard these tendencies as endearing rather than reprehensible. And this, of course, would simply make his position even more suggestive of one which promotes thoroughgoing rationality. The largest problem here seems to be that Dennett's proposal threatens to subsume any fairly broad-minded theory under the umbrella of theories like his - ie, theories which insist that people are essentially rational - and do this in a manner smelling suspiciously of aprioristic fiat, of argument by redefinition of terms. Is this *really* what 'rational' means; and, if so, can any bounds be set to what can be approved of?

The answer, perhaps, is that Dennett is right, so long as 'approval' is not too widely interpreted. What it should indicate is a high rating on some fairly specific scale, such as (what Dennett himself implicitly suggests at many places) a scale of survival value, or (just slightly less vacuously) of fit with the environment, in the sense of adaptedness to exploiting the opportunities and avoiding the dangers which it presents. (This is not, of course, going to be in general very easy to assess.) It then *is*, perhaps, necessarily true that any optimally efficient system is essentially rational.

6.2 Rationality and Normative Logic

The point of the foregoing has been to defend the view that the interpretability of behaviour entails its fundamental rationality. The notion of 'rationality' at issue seems no less vague for Follesdal's endeavours, nor Dennett's, but the main question was what it has to do with logic. My answer to this turns on the idea that the connection can be seen as mediated by the postulated cognitive theory cast in computational terms. The suggestion has already been made that algorithm-explainable phenomena are those susceptible of 'reason explanation'; and it has also to be recognised that in an important sense these are inter-determining. Insofar as the algorithm is seen to embody *interpreted* items of some kind, we have the argument of Dennett (1981a - see ch. 3 above), among others, that the interpretation proceeds necessarily from a prior intentional characterisation of the whole system; and we can now observe also that this characterisation essentially amounts to a reason-explanation of its behaviour.

An analogy was used in chapter 4, whereby the problem of determining the competence theory for a particular behavioural domain was likened to that of finding the best-fitting curve through a set of points on a graph. One has the observed behaviour, and one has to decide what the nature of the ideal is. It seems clear that, for instance in a reasoning task, the competence theory chosen specifies a function from premises to conclusions regarded as 'correct'. What we can now notice, is the interplay between the 'logic' thus defined, and the notion of rationality inherent in the overall intentional characterisation which grounds the interpretation of premises and conclusions. It looks as though we want to say, here, that the latter notion is somehow prior to the logic defined - which would be consistent with, say, the views of L. J. Cohen, and also those of Dennett. But now, although this notion of rationality is still obscure, and perhaps incorrigibly so, we have seen enough to note that it is somewhat loose, and related to an evaluation of, in some sense, the utility of behaviour in certain circumstances. This remark echoes a particular strain in the above discussion of Thagard's work.

Thagard tells us that a major component in the evaluation of a set of normative principles is its 'efficacy', which means 'the extent to which the principles and practices of a system lead to satisfaction of the relevant inferential goals'. This, on his account, has to be essentially whence the normativity of a system is derived: one *ought* to have an efficacious system. And, as observed before, the question now arises as to what these goals are, and whether they themselves are subject to criticism in point of their rationality. An example of such a goal might be the preservation of truth in inference, and indeed it seems clear that this is one which Thagard has in mind. But it is just this kind of goal which Dennett is suggesting, in the passage quoted above, might be inappropriate in certain circumstances - eg. if the costs of *ensuring* truth-preservation outweighed its benefits in

those circumstances. That is to say that, in those particular circumstances, a system which effectively promoted a certain inferential goal would still not be recommended because this would fail to coincide with the current best interests of the reasoner.

The lesson to be drawn here, in my view, is the same as previously noted - that the set of goals used in assessing the efficacy of a system has to be relativised to the circumstances of the reasoner. The same reasoner in a different situation, or a different reasoner in the same situation, might be served differently by a particular choice of inferential goals, and to that extent it might be rational to adopt a different set of normative principles. If we can suppose, what is enormously simplifying if not necessarily well furnished with empirical support, that individual people are similar in their cognitive apparatus etc., this yet leaves us with the strong suggestion that the inferential situation is crucial. Cohen might be right, for instance, to suggest that gamblers are well advised to stick to their fallacy. Except, of course, that this would then be the wrong way to put it, since what might well be a *statistician's* fallacy is revealed to be entirely justified for the gambler.

But it should be noted that we still are not in Cohen's apparent position of supposing something to be normatively vindicated simply because reasoners do it, in whatever situation they happen to be in. Thagard tackles Cohen about this issue, on the ground that he takes too strictly his analogy between the psychology/logic case, and the language/grammar case. In the former, but not the latter, we have normative force:

> Practice can be *improved*. Logical practice has improved enormously with the developments in deductive, inductive, and practical logic of the past several hundred years. There is no analogous sense in which linguists aim to improve the overall grammar of a linguistic population . . . (Thagard 1982, *op. cit.*, 35).

This is a point made before (cf. Rescher 1977 242*ff.*), and it seems to me a dubious one, if it is meant to allude to some essential, immutable difference between these cases. There are two problems. Firstly, Thagard has again confused the practices of logicians *qua* logicians (which no doubt have improved even as he says) with those of everyday reasoners (who, for all any evidence I have seen suggests, perform as badly as ever); moreover, logicians' practices are, after all, often theory-constructive, and it is arguably in this area that they have mostly improved. (Was even Aristotle's *inferential* practice all that poor? Surely he was primitive mainly in theory-construction - cf. Thagard and Nisbett, 1983: 'Modern symbolic logic provides methods for assessing the validity of complex arguments which are much more powerful than Aristotelian methods, which were largely restricted to the syllogism' (259) - this provides no ground for supposing that many of his actual arguments were invalid.) However, linguists' theory-constructive practices have surely also improved greatly, especially since Chomsky.

Secondly, although it's true enough that logicians *aim* to reform and improve, and that linguists (nowadays) do not, it is unclear that this is more than in some sense accidental: Thagard goes on to say

> The logician . . . is concerned to develop a set of principles which is inferentially optimal given the cognitive limitations of reasoners (Thagard, *loc. cit.*)

- and this is clearly so (if hardly expressed in terms the average logician might use), but could it not have been, equally, that the linguist strove to develop a set of communicatively optimal principles? Or *stylistically* optimal principles? Or whatever; might he not have had some goal, at any rate, which would render his activities similarly prescriptive and normative? If he had, though, his methods might have to change; and this is the point of the objection. No longer could he maintain the *laissez-faire* position of merely describing the practices of his subjects or informants (or himself). Such is what Thagard would maintain. The linguist would then have to adopt a wide reflective equilibrium approach, and import background psychological and other theories so as to allow the assessment of practices with respect to 'efficacy' and 'optimality'.

Cohen would disagree, of course. The linguist would have to compare his emerging theory with the intuitions of his informants: he could not characterise them as 'ungrammatical' unless he could show that his theory both conformed to and captured their intuitions (otherwise it would not be a theory of *their* practice), and also revealed those intuitions as being in 'contradiction' with it; and this would be impossible. But here we do seem to see a certain poverty in Cohen's theory. No doubt it's true that 'unless we assume appropriate intuitions to be correct, we cannot take the normative theory . . . that they support to be correct' (Cohen, *op. cit.* 319), but then perhaps the theory based on them *is not* 'correct'; which is to say, we might take something else as the standard of correctness - eg, efficacy - and compare the theory raised on intuitions with the theory raised on that.

This discussion of the linguistics analogy has been of value, in that it has brought to our attention an evident gap between competence theories and normative theories; a difference not recognised by Cohen. Competence theories - most obviously as one encounters them in linguistics, but elsewhere too - are things wrought solely for the purposes of *explanatory* theory, and should not in themselves be taken to have any normative import. Of course, they will entail a certain regularisation of the data - this happens even in linguistics, where the occasionally espoused ideal of describing multifarious 'idiolectal' grammars rarely seems to have any value, and is usually forsaken for something more general. This in itself provides no opportunity for accusations of error. The creation of a normative theory needs additional considerations, such as Thagard suggests, with the competence theory having an important *constraining* role to play insofar as it can show what things are possible for the reasoner. What the reasoner cannot do (or avoid), he cannot be

normatively required to do (or avoid); but equally, he need not be required to do or to avoid something, simply because he is able. Typically, of course, normative theories, at least in logic, do not require a particular (or any) inference, but simply allow any of several while forbidding all others. This is more than one is going to get out of a competence theory derived from observations and data, since many permissible inferences are just never made (*cf.* the discussion in chapter 4, particularly with reference to Pylyshyn, 1973).

* * *

The discussion of the last few paragraphs in some respects parallels one offered more recently by Stich (1985), who also discusses Cohen's position in the light of the linguistics analogy. Stich concludes that Cohen is committed to the idea that there are many normative theories of cognition and reasoning. But then, he asks, *which is the right one*? Cohen might respond to this with a 'thoroughgoing relativism: my normative theory is the right one *for me*, yours is the right one *for you*', but Stich rejects the idea as too unpalatable:

> We are not in the least inclined to say that any old inference is normatively acceptable for a subject merely because it accords with the rules which constitute his cognitive competence. If the inference is stupid or irrational, and if it accords with the subject's cognitive competence, then his competence is stupid or irrational too, in this quarter at least. (132)

After this, he returns to the idea of normative correctness as expert RE, much as in Stich and Nisbett (1983), *op. cit.*

My response here is that for one thing, Stich has underestimated the role of the normative theory in the initial formulation of an account of competence, but that for another he has artificially limited the range of options open to normative theorising at this point. For he excludes my option, which is to suppose that competences extend across *subjects*, but not necessarily across *tasks*, or more accurately task domains. I thus embrace a form of relativism; but I resist 'the unhappy conclusion that a patently irrational inferential strategy might turn out to be the normatively correct one', because I insist that the relationships here are more subtle than Stich makes out. In circumstances where strategies of this kind seem to arise, pressure builds up on the components of the description of the situation. If the subject continues to be seen as irrational, the possibility of coherently modelling his cognitive processes is eroded. If he does not, the characterisation of the inference as normatively incorrect cannot be sustained. In either case, the competence theory behind the cognitive account of the strategy is forced to respond. Such circumstances are inherently unstable; profitable theorising demands that they gravitate towards a resolution.

6.3 Concluding Remarks

We have encountered the idea that normative principles are to be relativised to the circumstances of the subject. Likewise seems the case of competence theories. In different situations, an ideal reasoner, we have supposed, will behave differently, and properly so. What we want to say about the relation of competence theories to normative theories is then really that prior to both of them is that ill-defined notion, rationality of action in a given type of situation. It might be said that the elaboration of these two kinds of theory constitutes two different ways of trying to define that notion. But really it's always curiously circular because, to start with, the implicit idea of rationality is used in perceiving behaviour as intelligible at all, and the resulting interpretation strongly conditions the nature of any cognitive theory of that behaviour; but then also our theory of what is normatively correct conditions the selection by such a cognitive theory of certain aspects of behaviour as 'competent'; but that normative theory depends for a crucial part of its justification on the sort of cognitive account of human abilities that it is being used to help found.

If we are lucky, the circle is really an upward spiral. In each new loop, as it were, the notions of competence and normativity used are those from the theories in the previous level, 'bootstrapping' themselves aloft. Thus we might support the idea of some kind of mutual progress in the concurrent development of the two types of theory. We might hold that in some sense it would be better to see this as the construction of a single, unified theory, in which the notions of competence and normativity gradually converge to meet in unity with the rational ideal. In any event, the notion of 'rationality' itself always remains logically pre-theoretic, a springboard for the process, and an ultimate arbiter should it threaten to get out of hand. If our account is right, however, it is also an arbiter, or (to change the metaphor) a referee, liable at any time to move the goalposts in unexpected ways, and indeed to present goalposts in different places when the game is played in differing circumstances. This state of affairs should not necessarily occasion any great surprise. As Wittgenstein indicated, the way the game is played is the ultimate fact of the matter, and there's no guarantee that the game is played always in the same way; it may consist of any number of relatively inconsistent subgames.

Part II

7

Introduction to Part II: Psychology and Syllogisms

This second part of the thesis consists essentially of a detailed examination of P. N. Johnson-Laird's theory of internal representation and inference. In the process, we shall be paying particularly close attention to his account of syllogistic reasoning. Our eventual purpose is to expose some of the important presuppositions of computational cognitive psychology, as it applies to reasoning. Having spent some time in a preliminary exploration of background material, we are in a position to look in a detailed way at an example of an actual piece of psychological theory, in order to see how these things show up. One of the questions that will interest us here, is the way in which theory is related to experimental test, and we shall therefore devote significant effort to a close inspection of important experimental data, and the claims surrounding it. The task of the present introduction is to provide some motivation for looking at the domain of syllogisms, and this theory in particular.

As noted before, it will not be necessary to cover in detail the history of investigations into the syllogistic task, since this has been done for us in a most competent fashion, by Evans (1982) and Johnson-Laird himself (1983). We shall eventually come to consider some of Johnson-Laird's criticisms of the previous literature, but for now we may be content to pass over this history relatively briefly.

The syllogistic task is of general interest, firstly because it involves quantified propositions. Earlier (in chapter 2) we looked at Johnson-Laird's (1975) theory of propositional reasoning, and noted that he regarded quantified reasoning as quite a different kettle of fish, calling for the 'combination of information' in 'mental models', rather than the application of some kind of rule schemata. He persists in this view, although he later extends the remit of a mental-model-based account to cover propositional reasoning as well. Certainly, from a purely formal point of view, quantification introduces a great many extra complexities into a logical system, that can cause trouble for inference schemata. From a

psychological point of view, there is no doubt that many cases of quantified reasoning are substantially more difficult than most propositional tasks, and that serious processing issues are thereby raised.

There are various formal approaches to the problem of extending propositional logic to operate with quantified propositions. In natural deduction systems, one tackles the matter in basically the same way as for the logical connectives, by postulating introduction and elimination rules for the quantifiers, which allow, for instance, reasoning about *all* objects to go ahead through reasoning propositionally about *one* object, on the basis that that one object is understood to be *arbitrary*, and in some sense therefore *representative* of all other objects, insofar as it differs from them in no respect relevant to the current argument. In a sense, this is perhaps at the root of Johnson-Laird's intuitive notion of a mental model. For he suggests that such a model represents a class of entities by containing some arbitrary number of arbitrary exemplars of that class. Reasoning about these entities generalises precisely because of their arbitrariness. (At least for non-numerical quantification, it is generally unnecessary for this number to be greater than one, but Johnson-Laird holds that, if nothing else, some number greater than one has more intuitive psychological appeal.) With such a model, we can reason about the items it contains, and then implicitly generalise the results on the grounds of arbitrariness. If one formalises syllogisms in a sorted first-order logic, it becomes quite natural, as we shall see, to regard Johnson-Laird's models as rather similar to sets of statements about the sorted objects.

This suggestion is in fact somewhat radical in terms of Johnson-Laird's theory, since it is one of his more important claims that his models represent whatever they do represent precisely *not* as anything like a set of propositional statements, but rather by somehow *analogically* representing its structure. He is here reflecting a distinction which has been important in several departments of cognitive psychology, and the issue is one that will crop up frequently in the following chapters. One of our claims will be that he has overestimated the extent to which any such distinction with respect to his own models can be supported either on formal or on empirical grounds.

Such claims detract little from the interest of his theory, of course, which remains one of the most provocative to appear in cognitive science. It is an interesting theory, particularly, in that it tries to say something highly detailed psychologically (about the cognitive processes involved in solving syllogisms), but at the same time very principled in its logical aspect (so far as it tries to capture the sense in which this is a process of *reasoning*, based on *rational* operations). A significant observation concerning preceding work on syllogistic reasoning is the lack, in many cases, of either of these features, and certainly the almost complete absence of any attempt at their combination. Early theorising about syllogisms was in the tradition of verbal testing, and more emphasis was given to the statistical

treatment of the experimental data than to the question about what sort of *processes* might have led to their production. The closest one gets to a theory about any kind of mechanism is the so-called "atmosphere effect" story (Woodworth and Sells, 1935), which postulated that reasoners were simply somehow affected by the 'mood' (partly in the Aristotelian sense, but also with vaguer overtones) of the syllogism, so that for instance a generally universal or particular mood was liable to lead to a universal or particular conclusion (or agreement with one if offered it), regardless of its validity or otherwise. This turned out to be a moderately successful predictor of experimental data, but it obviously says nothing about either of the points noted above.

Atmosphere theory was succeeded by the 'illicit conversion' theory of Chapman and Chapman (1959), where it is supposed that subjects convert premises, for instance those of the form *All A are B* into (more or less) *All and only A are B*, prior to working with them thereafter in a wholly logical fashion. Here we have an attempt to say more about both processing and the role of logic, but only a rather vague one. Its most important feature, perhaps, is the insistence that subjects are to be seen as reasoning *logically*, in accordance with normatively standard rules of inference, once they have unreasonably converted their premises. Evans (1982) goes to some lengths to stress that there is actually very little empirical evidence to suggest that subjects *do* reason in a wholly logical way. He regards this as a prejudice which is held by many psychologists, more in the face of the evidence than anything else, out of a conviction that man must be essentially a rational animal. We have already addressed several of the main issues surrounding this point, in Part I, where we noted that Johnson-Laird is himself particularly prey to the prejudice in question. We concluded that such a prejudice is in some respects justified, but not if the logic required by the theory is assumed *a priori* to be (say) first-order classical logic, or indeed Aristotelian syllogistic. The approach is particularly necessary if one's account of reasoning is to be a cognitive one, offering a computational model of processing. The Chapmans' account is of course not of that kind, and to that extent its importance to the theme of this thesis is diminished. But there is value in highlighting the fact that where a processing theory assumes a particular logic, a salient possibility for dealing with recalcitrant data is to suppose that subjects do 'illicitly convert', or in some other way misrepresent or misunderstand the premises of their inferences.

Johnson-Laird's theory, however, postulates that subjects understand the premises perfectly well. On the other hand, they may get into trouble when it comes to putting them together to form a syllogism, which is crucial given that the whole problem of solving a syllogism can be seen as the problem of deciding what the combination of the premises means, and thus arriving at its implications. Johnson-Laird's theory is among very few that suggest an *algorithm* whereby subjects are supposed to compute their conclusions, and it is largely because it does so that Johnson-Laird is led to make some very strong

claims about the implications of his theory for our assessment of the rationality, and the reasoning competence, of our subjects, in the syllogistic domain. Johnson-Laird suggests that this act of combining the premises in many cases is one fraught with difficulty and one that imposes such demands on cognitive resources that it may overtax restricted mechanisms such as short-term memory. A processing system which in principle is capable of running until it generates a correct conclusion every time may thus be thwarted and driven into error. This, of course, describes a system very much in the mould we considered in Part I, where the 'central algorithm' provides the basis of perfect competence for the task.

The domain of syllogisms commands attention, perhaps most centrally for just the reason that it focuses on the issue of interpretation and understanding of premises. It is because of this feature that Johnson-Laird can make it the cornerstone of his mental-model theory, which purports to extend its coverage to virtually all significant cognitive phenomena. It is because of this, too, that Johnson-Laird can summon some plausibility for setting up the explanation of an apparently syllogistic reasoning performance as his crucial test case for cognitive science (see Chapter 2, above). It is also because of this that what seem on the face of things to be a particularly boring range of obsolete logical curios, can be seen as being actually the simplest, and hence most investigable, of the really interesting problems of cognitive psychology. It needs to be noted, however, that this interest arises almost wholly from syllogisms' potential wider relevance, and that a theory which addresses only syllogisms, and cannot be shown to extend insights to other areas, is all but worthless, except on the doubtful view that syllogistic forms are themselves a ubiquitous feature of everyday discourse, and that they are there processed in the same way as in laboratory tasks. Evaluating claims for the extendibility of Johnson-Laird's theory is accordingly one of our more important aims.

Our strategy will be to begin by expounding Johnson-Laird's theory in some detail, taking into account its interesting development since its first appearance and the way it grows out of empirical studies, and encompassing some wider aspects of the idea of mental models in areas other than syllogisms specifically. Then, we shall pay considerable attention to the 'central algorithm' that it embodies, directing ourselves in particular to the question of its status as an 'effective procedure' for solving syllogisms, and the consequences of its being one if it is. We shall examine how far we can uphold Johnson-Laird's crucial claims that his system is importantly different from various other theories that preceded it, and that its processing model shows the nature of the internal representation used (ie, the mental models) to be irreducibly *semantic* and *analogical*. Our contention will be that these aspects of the theory have been overstated in important ways. In conclusion, we will look also at Johnson-Laird's claims about the necessity of casting a theory in the form of an effective procedure in order for it to be explanatory, and we will assess the explanatory achievements of his account of reasoning.

We now proceed, therefore, to examine the actual mechanism Johnson-Laird proposes, with the aim of discovering its detailed motivation - the reasons for its being the way it is - and subsequently to view this both on its merits as claimed for it, and in the light of the preceding discussions.

8

Johnson-Laird's Theory of Syllogistic Reasoning: An Introduction

This chapter takes the form essentially of a commentary, in more or less the old-fashioned philosophical sense of a critical exposition, on Johnson-Laird's theory of syllogistic reasoning. This is undertaken not only for its own interest (which however is considerable), but as necessary groundwork for the chapters that follow. It also seems natural in the context of treating it as a kind of case study on an approach to investigating reasoning. My strategy will be to expound the two main versions of the theory along with their associated empirical investigations, in parallel, so that we can compare them and observe the salient points of difference. In the process, we shall discuss some of the issues which arise about experimental design and interpretation, and generally the more psychologically-oriented aspects of the subject, including basic questions concerning the nature of internal representation. In later chapters, we shall devote more sustained attention to the detailed representational, logical and semantic claims which Johnson-Laird regards as so important in his work, where we shall widen our view to take in areas outside the specific topic of syllogistic reasoning.

Johnson-Laird's syllogistic theory purports in fact to be a theory of reasoning with quantifiers in general, but as we shall see, there are good reasons to doubt that it can be extended very far in this direction, and in fact it alludes only peripherally to the very possibility. It began as a theory mainly of syllogistic reasoning, presented in 'Models of Deduction' (1975). In this paper, Johnson-Laird starts out by developing the account of propositional reasoning that we looked at in chapter 2, above. He then addresses the question 'whether it is possible to devise a general model of inference with quantifiers along the lines of the model for propositional inference' (38). The model which then follows is the prototype for all his later accounts (although it bears little apparent resemblance to the model for propositional inference).

The model is developed as an explanation for certain data collected in a set of experiments, which turn out to be the ones reported in Johnson-Laird and Steedman (1978), where the theory receives further elaboration into what it seems fair to regard as its first fully worked-out form. Subsequently, this model undergoes several changes in detail, some of which, taken together with the explanations offered of their motivation, are quite revealing about the underpinnings of the whole project.

8.1 General Outline

Johnson-Laird's model, in both its versions, is based on two essential notions, which both appeared in his 1975 paper, and which now form the basis of a much wider-ranging theory. One is that reasoners proceed by constructing 'mental models' which represent, in some sense, what is asserted in the premises; the other is that there are various ways of doing this, not all of which may be correct, and that what drives the inferential mechanism is the fact that an inference is valid if and only if the conclusion is true in all possible correct representations of the premises (all circumstances in which the premises are true).

His beliefs about the ways in which these essentials are 'implemented' in the human reasoner, have changed through time, but can be broken into two basic sets:
(a) those forming the theory of Johnson-Laird and Steedman 1978;
(b) those forming the theory of Johnson-Laird and Bara 1983.
(There is a version in Johnson-Laird, 1983, very similar to (b), which we shall mention only occasionally in this chapter.) There are some fairly substantial differences of detail between these, but they both involve the following basic steps. (*Note*: For the sake of brevity, I tend in what follows to attribute all quotations and paraphrases, from both papers, to Johnson-Laird alone. This is not intended to belittle the contributions of his collaborators.)

(1) *The premises are represented, together, as a combined mental model.* Traditional scholastic logic recognises 4 forms of proposition, designated by mnemonic letters:

All X are Y	(A)
Some X are Y	(I)
No X are Y	(E)
Some X are not Y	(O)

Johnson-Laird uses these, as well as his own set of 'figures',

1.	2.	3.	4.
A B	B A	A B	B A
B C	C B	C B	B C

which allow any pair of premises to be specified, as eg. the syllogism

All A are B
Some B are C

is '1AI'. (Also, I will at times refer to a proposition using the form-mnemonic in lower-case between the two terms, eg 'All A are B' appears as 'AaB', 'Some B are A' as 'BiA'.) There is a fairly trivial algorithm which translates premises of these forms into models as follows:

All A are B	——→	a = b (b)
Some A are B	——→	a = b (a) (b)
No A are B	——→	a ≠ b a ≠ b
Some A are not B	——→	a ≠ b (a) = b

(there are important notational differences between this and (a) and (b), but we neglect these for simplicity just at present).

The idea is that the tokens in the models represent individuals in the sets (eg. of As) being modelled, and that any arbitrary number of these can be regarded as modelling the entire actual set. Items in brackets are supposed to be 'optional', in the sense that the model is uncommitted as to their actual existence: hence, eg, in the case of the A proposition there may or may not be Bs which are not As. (Johnson-Laird says that to obtain a Boolean interpretation of universal quantification, with no existential import, one can simply make all the A terms optional - see (b), 38.) The identity links in the model simply indicate that connected tokens 'refer', in some sense, to the same real world individual (or perhaps I should say set of arbitrary real world individuals), which is asserted to belong to both sets mentioned in the proposition. Links of denied identity show the existence of individuals which *do not* belong to both sets. We shall return in subsequent chapters to the many problematical aspects of this suggestion.

Both (a) and (b) assume that the first premise is represented with one of these models, and that the information contained in the second is then added in some way. There is a question as to whether this proceeds by separately modelling both premises and combining the results, or by 'adding in' the second premise's information to the model of the first. In (a) the answer to this is left obscure, but in (b) it is held to be immaterial (40). However, a crucial point is the importance of the 'middle term'. Suppose we have the

syllogism (1AI):

All A are B
Some B are C

then the model is (initially)

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

This result is due to the assumption that model-construction is guided by a *heuristic*, which is essentially the same in both versions of the theory:

(a) 'there is a heuristic bias toward forming thoroughgoing connections between all the classes, that is, a bias toward linking up end items by way of middle items' (78)

(b) '. . . reasoners are guided by the heuristic of trying to maximise the greatest number of different roles on the fewest number of individuals' (46).

Of course, this heuristic is not the logically 'best' one, and hence leads to a characteristic class of possible errors, as will be discussed. The arguments in its favour depend largely on the ubiquitous dogma of 'parsimony', and the fact that it seems to fit the experimental data.

(2) *A putative conclusion is formulated on the basis of the model.* Given some combined representation as described, the task of the reasoner is to draw from it a conclusion to the syllogism. This must link the 'end items' (from the sets A and C), and for preference should be in the correct propositional form - there is a simple algorithm available for doing this, which is exploited in both versions of the theory. (It has a slightly different form in (b) due to the different notational devices used in that paper; but it's essentially equivalent.) The rules are: where a 'path' is a series of 'positive' (identity) or 'negative' (non-identity with identity or null) links from an A token to a C token, if the model contains (i) at least one negative path, then the conclusion is of form O, unless (ii) there are only negative paths, when it is E. If there is at least one positive path, the conclusion is I unless all paths are positive, when it is A. If there are only indeterminate (null or positive with null) paths, there is no valid conclusion of the required sort. It's *assumed* (there's no model-based explanation for it) that people naturally try to establish relationships between the end terms here, and that thus 'the middle term tends not to be referred to . . .' ((b), 46). (This tendency is certain to be influenced by the instruction 'to restrict their answers to one of the four moods or else to state that no valid conclusion followed from the premises' ((a), 71), since an earlier experiment without the instruction *had* produced a few responses involving the middle term; we are not told of such an instruction in (b), but the

nature of the data makes it hard to doubt that there was one.)

Since the representation initially constructed (on the basis of the heuristic) is not necessarily optimal, the conclusion 'read off' in this fashion may not, in fact, be valid. It is valid only if it can be read off any possible representation of the premises. Johnson-Laird's models have the useful property of, themselves, each representing a large number of possible 'states of affairs' in which the premises are true (by virtue of the items in brackets); but not always all of these, and so alternatives must be sought.

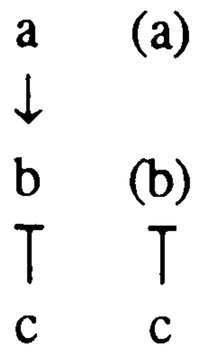
(3) *The conclusion is 'tested' against other possible models.* In this phase of the procedure, the differences between (a) and (b) become greater, but the essential process is the progressive deformation of the initial model, in various ways, it always being understood that the model *must remain consistent with the premises.* Links can be broken, new links established, items added or removed, so long as this cardinal principle is observed. At each stage a new conclusion is read off: in (a), the system is so arranged that further deformation (by the procedures provided) rapidly becomes impossible, and the final model always yields the correct conclusion (be it only that there is no *valid* conclusion); in (b), one has to keep track of the intermediate conclusions, the correct one being any which is valid in all the stages, else 'no valid conclusion'. It turns out in both cases that the number of deforming steps a model can undergo before it reaches a stage where further modification by the rules is impossible, is very small. There are *at most three* of these 'alternative models' to be considered.

8.2 Some Theoretical Details

(a) The theory of Johnson-Laird and Steedman, 1978, is highly precise and constrained. Models are formed with *arrows* (or 'stopped arrows') as the links, which are supposed to indicate a directional bias in reading off conclusions (this is to explain the 'figural effect' in the experimental data, to which we return later). Thus we have, eg, for 1AA:

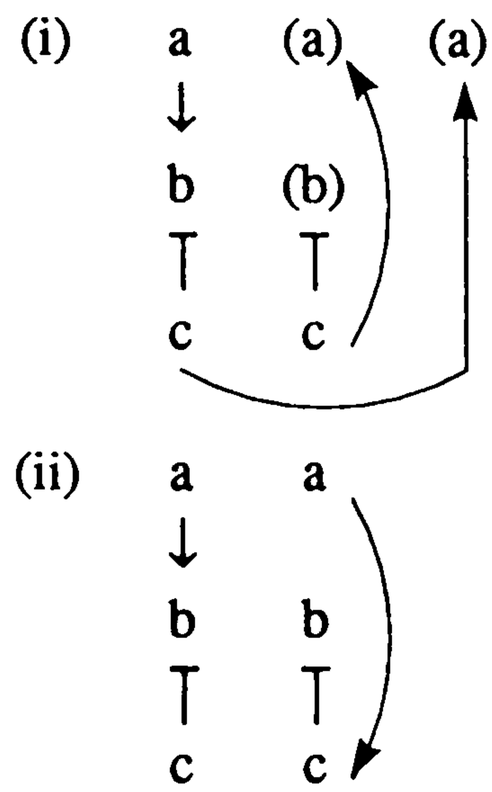
```
a     a
↓     ↓
b     b     (b)
↓     ↓     ↓
c     c     c     (c)
```

and for 3IE:



The model-construction procedure always constructs the same model, in all details, given the same premises, so we have a clear set of 'initial representation conclusions' (IRCs) for each syllogism, as determined by the algorithm outlined above. Testing then proceeds by attempting either to break positive links, or to establish new ones past existing negative links, if this can be done consistently.

Some cases are interesting: consider, eg, 3IE (depicted above) which initially suggests 'No A are C' or 'No C are A'. This can be tested in two ways, one equivalent to testing the first conclusion mentioned, the other to testing the other:



(*cf.* Johnson-Laird and Steedman 1978 81,94). This shows that, after testing, the conclusion-forming algorithm is abandoned, since there are paths in (i) which clearly suggest an A conclusion. What happens is that if the IRC is impugned by testing, the correct conclusion is 'no valid conclusion' unless further or other testing establishes one: hence only the testing shown as (ii) leads to the correct, valid conclusion in this case. What's important is that the procedure has 'forked' along two distinct paths, and there is no route back once one has been taken. Which path is chosen depends, apparently, on which way the IRC is read off the model (A-C or C-A), although this is not explained in any detail (see Johnson-Laird and Steedman 1978, 94).

Detailed exposition of these processes for each of the 64 possible syllogistic forms produces a set of conclusions within which it is predicted that subjects' responses will fall, in the experimental task. Agreement with the data is in fact remarkable, and will be considered later.

* * *

(b) The more recent (and allegedly 'final') version of the theory is substantially different, at least in emphasis. Now, instead of having one model which is gradually altered, we have an account in terms of up to three separate models which can be constructed for any pair of premises. A slightly earlier statement (Johnson-Laird 1983, ch. 5) leaves it obscure as to whether these must be constructed in a determinate *order* (although it strongly suggests so, and indicates that they *could* be derived from one initial model by stepwise distortions), but the final form has it that, as we have seen, there is a fixed heuristic guiding construction of the initial model, and that the testing process 'is based on an initial model' (49). This process also consists in the attempt 'either to break positive identities or to bypass negative barriers' (*loc. cit.*): it is thus highly similar in actual effect to the procedure in (a).

The models are in this account notationally different, being set out as:

A:

a = b
a = b
0b

I:

a = b
0a 0b

E:

a
a
———
b
b

O:

a
a
———
0a b
b

where for any token '0t' represents an optional token (as the brackets in (a)), and the line represents the impossibility of forming paths between things it separates, analogously to the stopped arrows. If one of these 'negative barriers' has tokens of the same kind (optional or not) on either side, then it is said to be 'penetrable', eg. as in the O case illustrated. The conclusion-forming algorithm exploits this: if there are not both As and Cs on the same side of a barrier, the conclusion is O if the barrier is penetrable, else E.

Testing proceeds via 5 procedures, variously applicable:

(i) Breaking of positive links

(ii) Shifting of optional tokens of end terms round negative barriers

(iii) Adding more optional end-tokens of already existing types

(iv) Swapping optional and non-optional middle terms round a barrier

(iv) Where there are 2 barriers, shifting tokens of one end term round to the same side as those of the other.

These procedures yield at most 3 models for any syllogism, and conclusions are read off these by the same algorithm. As already remarked, a valid conclusion is one derivable from *all* the models (but note here that, eg, an O conclusion is taken as valid in a model where the algorithm yields an E conclusion - the Lisp model does not implement this part of the procedure). The number of models predicted for each syllogism is what is said to be the crucial result here (Johnson-Laird and Bara 'obviously . . . do not wish to defend the psychological reality of these procedures' (52)): it predicts that the more there are, the more difficult the problem will be, given certain assumptions about memory and processing, some of which are considered in the following section.

8.3 The Experimental Studies

The two main studies I shall consider are those described in (a) Johnson-Laird and Steedman 1978, experiment 2, second test, and (b) Johnson-Laird and Bara 1983, experiment 3. These experiments are supposed to be very similar; indeed (b) seems to be a straightforward replication of (a). The results are rather different in various ways, but let us first consider how well each set of data corresponds with its accompanying theory.

(a) The experiment reported in Johnson-Laird and Steedman 1978 (actually conducted in America by Janellen Huttenlocher, or one of her students) involved 20 students being asked to provide conclusions to all 64 possible pairs of syllogistic premises, constructed with different 'sensible contents', presented one after the other on cards. The subjects' response latencies were timed by the experimenter (presumably, with a stopwatch). The subjects were told to frame their answers in correct propositional form, and apparently did so with few if any exceptions. The results are given as a set of tables in which each pair of premises determines a cell where there appear the computer-generated conclusions predicted by the theory, and the number of subjects (out of 20) offering each of these. A few unpredicted conclusions are mentioned, if these were given by at least two subjects either on the test in question, or on the previous one conducted almost identically on the same subjects a week earlier. The mean latency for the *correct* conclusion in each case is also given.

The correspondence between the predictions and the results is in general astonishing; 95% of all responses are within the predicted range. The theory also predicts certain 'figural effects', the nature of which, as noted, is supposed to be captured by the arrows in the models. These arrows all point the same way in the first two of Johnson-Laird's syllogistic figures, but not in the last two; this, he says, creates a bias towards a certain form

of conclusion, *viz.* A-C in the first figure, and C-A in the second. Hence 1AA (illustrated above) gives us 'All A are C', preferentially over 'All C are A'. One finds, indeed, that the 'antifigural' responses are rarely given, and Johnson-Laird argues persuasively that this helps to account for the difficulty (measured as rarity of correct responses) of premises such as 1EI and 2IE, where the correct conclusion is antifigural (72).

Johnson-Laird also has another major prediction, which is that where his theory requires that a model be tested before the correct conclusion is reached, this will increase the difficulty of the problem. The more testing that is required, the more difficult (and time-consuming) the problem should be. He says: '...80.4% of responses to problems where a test leads to no modification were correct, whereas only 46.5% of responses where a test leads to a modified conclusion were correct; the difference was apparent in the results of all 20 subjects.' I confess myself unable to find in his data any basis for the figure he gives of 46.5% (although the other seems correct), but in any event I claim that this neglects a notable phenomenon in the data where the premises contain no A proposition (ie 36 out of the 64 pairs). If there is no A premise, the conclusion invariably requires testing, is usually 'no valid conclusion' (unless the premises are EI or IE), and is obtained by 81% of the subjects! Moreover, it is often obtained very swiftly; Johnson-Laird himself remarks that 'those premises which yielded many correct answers also yielded them rapidly' (76). This indicates that A premises are somehow a source of difficulty in testing, and I have dubbed this the *A-effect* (see next section).

Johnson-Laird goes on (84) to produce some very detailed predictions about performance, which depend on factors such as the number of paths there are in the initial representation, and how hard it might be to break them during testing: I regard these as wholly speculative, and describe some of them in the next section. Worth mentioning, though, is Johnson-Laird's final suggestion that the apparent ease of certain syllogisms with two negative premises (eg. 1OO) might be due to some subjects' having 'learned to interpret two negative links in a path as indeterminate' - the theory itself giving no clue as to how such learning might take place.

Little of value can be said about the latency figures provided. Even Johnson-Laird claims nothing except the quoted remark, and this is not, itself, particularly supportive of his theory, since a large number of the correct answers were the result of copious testing (*cf.* below). But questions ought to be asked, in any event, about how these figures were measured (at what points timing began and stopped, etc.) since the data are so incredible; to ask us to believe that the average subject solves the 1II syllogism (correctly!) in 1.9 seconds (a full second faster than the 1AA) is far-fetched, and anyway adds nothing in plausibility to Johnson-Laird's theory's assertion that this result requires a testing step (whereas the 1AA case does not). Indeed, if they could be taken seriously, some of these

data would cast serious doubt on Johnson-Laird's whole model - but in fact I think no-one should be tempted to attach any importance to this.

* * *

(b) Johnson-Laird and Bara (1983) contains reports of three experiments. The first of these is basically a replication of that just discussed, with the important difference that the subjects were only given 10 seconds in which to formulate their responses. The main result of this was a large increase in the number of incorrect 'no valid conclusion' responses observed, *increasing over the 4 figures*, and Johnson-Laird felt obliged by this to abandon the theoretical model of (a) on the grounds that the new results indicated 'that certain figures make it difficult to construct a model in the first place' (29) and that this would offer a basis for a principled explanation of all the figural effects. ((a), of course, assumed no differences in difficulty in model-building.) The new theory accordingly states that the effects arise from 'the processes of integrating the information from the premises within working memory'. This indicates that the effects should arise in some form in asyllogistic cases, such as 3-term series problems. A second experiment demonstrates that this is indeed the case. It is thus seen to be necessary to formulate a new theory to account for the effects in syllogisms, in these terms.

The notion that restrictions on working memory cause difficulties, is embodied in the theory in two places, *viz.* the construction of the models and the deciding of the final conclusion. It is now supposed that when the information from the second premise has to be added to the model (or, as it might be, the two separate models have to be combined), this is facilitated if the models are oriented so that the occurrences of the tokens representing the 'middle term' of the syllogism are adjacent. (The idea comes from Hunter's (1957) story on 3-term series problems.) This is only naturally the case in the first figure; otherwise, various adjustments have to be made. In the second figure, the premises can simply be re-ordered (which is explained as holding the representation of the second premise in working memory while renewing that of the first) and in the 3rd and 4th figures, increasingly complex rearrangements are required. (These involve the 'switching round' of models, so that the tokens appear on the opposite side to their original one.) The result is that it becomes plausible to suppose that the task of model building (regarded as unproblematic in (a)) becomes more difficult through the figures. Interestingly, it is nowhere suggested that *negation* might have any effect on difficulty of modelling, despite obvious differences in the models, and well-known psycholinguistic problems in this area.

Certain syllogisms, as already noted, give rise to more models than others, with a maximum of three, and since a valid conclusion to such a syllogism has to be one valid in all these models any restriction which makes it difficult to 'examine' several models simultaneously (assuming this is in fact necessary - perhaps the intermediate conclusions are just

stored as strings, but Johnson-Laird doesn't consider this) will make for greater difficulty as the number of models increases. Hence the question of the abilities of working memory becomes crucial in Johnson-Laird's theory, and he claims experimental support (unpublished) in the form of a correlation between performance on syllogisms and on certain kinds of memory tests (64).

The prediction of the original figural effect, as was supposed to be explained by the arrows in (a), is now based on a further assumption - that working memory takes the form of a 'queue', (as opposed to a stack) so that whatever order things go in, is the order in which they tend to come out. Hence, if syllogistic terms go in in the order A-B-B-C, they'll tend to re-emerge in the same order (or at least those that emerge at all will; in this case A and C, in that order). This 'first in, first out principle' (41), despite its centrality, is given little elucidation, and no real independent empirical support. It's doubtful, too, how well it fits with Johnson-Laird's claim (1983 71-2) that syllogisms, with detectable figural effects, occur often in daily life, since it seems likely that the surface syntax hiding the underlying propositional form will often distort this ordering - *cf.* 'they also serve who only stand and wait' (an A proposition with term order reversed, *cf.* Cohen and Nagel, 1972 37) - and if a 'simple' syllogism should appear in a very complex form requiring much preprocessing, it is unclear whether the final canonical form derived will predict the form of the conclusion or not. The other essential assumption, ie. that information in memory tends to 'fade away' for some reason, is of course common in the literature.

The theory as it now appears thus makes certain predictions about the behaviour of subjects in the experimental task: some syllogisms should be more difficult than others, and there should be a general increase in the difficulty of otherwise similar syllogisms through the four figures. Johnson-Laird therefore proposes an experiment to test these predictions. This experiment (Experiment 3) is in all given details a straight replication of the one in (a) (with the exception noted above, that we are not told whether the subjects were asked to state their conclusions in normal propositional form). Since this is the case, I do not see why Johnson-Laird doesn't provide a reanalysis of the results obtained in (a), and point out whether or not the new theory's predictions are fulfilled there. Evidently, he thinks that a new experiment will provide further data, and that this will be more interesting, but the old data should show the same patterns nonetheless.

In the event, Johnson-Laird finds that his predictions are all borne out. The figural effect as demonstrated in (a) is again massively present, and there is a substantial decline in the number of correct valid conclusions as the number of models increases and through the figures (with no apparent interaction between the two) (56). The percentage of erroneous 'no valid conclusion' responses also increases through the figures and with larger numbers of models. Once again, there are problems with the latency data. These data are

in fact remarkable as compared with those from (a); we are not told of any difference in the manner of their collection, but they are generally much slower and more plausible. Since they are only given for *correct* conclusions, and these are comparatively rare, latency 'cannot be used as a general measure', except in the one-model problems. I am not convinced, however, that the latencies are reliable enough to prove anything even here.

An interesting further feature of Johnson-Laird's data emerges both in the first experiment of (b), and in the third. It is that 'where the conclusion is in the same mood as just one of the premises, the end term of the premise tends to play the same grammatical role in the conclusion as it did in the premise itself' (27, 55); eg, with the premises 3AI, if the subjects drew an I conclusion, it was likely to be CiA. What is mainly interesting about it is that Johnson-Laird apparently regards it as inexplicable according to his theory. Now the same effect occurs in the first two figures, of course, but there it is simply the original figural effect, as discussed in (a). (If this seems obscure, consider that the grammatical-role-based definition just given always leads to an A-C bias in the first figure, and a C-A bias in the second.) Allegiance to 'parsimony' would suggest that the explanation should be essentially similar for the phenomenon in all figures, but Johnson-Laird holds out no hope that his working-memory story will be able to do the job (*cf.* (b), 60). Hence, either there is a major peculiarity in syllogistic reasoning such that two parts of essentially the same phenomenon have different explanations, or Johnson-Laird's explanation of the figural effect is wrong. I suspect that further investigation of this matter would expose a serious problem for the whole theory.

Another interesting result is that in cases where there is an O premise one finds a number of conclusions of the I form. This is almost to be expected, given the way O propositions are interpreted in ordinary speech, but curiously it did not appear *at all* in the data for (a); in (b), however, the phenomenon accounts for 16.78% of all conclusions to premises of this kind. There is no obvious reason for this, and Johnson-Laird's model is constitutionally incapable of handling it (the conclusion-drawing algorithm cannot produce it): Johnson-Laird puts it down to some sort of 'Gricean implicature' (28). Indeed, Gricean principles do predict just the observed cases, and these go *against* the grammatical-role trend just discussed, in those syllogisms where I and O premises are combined.

8.4 The Effect of 'A' Premises in Syllogistic Reasoning

In this section, I want to look in a little more detail at a certain aspect of Johnson-Laird's experimental results. To see properly the way Johnson-Laird's theory grows out of his empirical work, and is inextricably bound up with it, it is worthwhile to probe further the way it fits with the data, and the way in which gaps can appear. I shall claim that there exists at least one important phenomenon unaccounted for by his theory, offer a

conjecture as to how his theory could easily be made to handle it, and then suggest that this is not altogether without significance from the point of view of sustaining some of his logical claims.

The phenomenon in question is best exposed by considering the role of the testing phase in the account of the solving of certain syllogisms. Most of the discussion here will assume the data found in Johnson-Laird and Steedman (1978 *op. cit.*), since that is where it emerges most clearly, and hence for convenience I shall refer to the version of the processing model found there, which is already described above. I shall also again refer mainly to the data from the second experiment, second test.

Take, for instance, the 1AI premise-pair: the 'initial representation' (IR) for this, on Johnson-Laird's theory, immediately suggests the conclusion 'Some A are C'. In fact, in this example, the suggested conclusion is invalid - there is no valid conclusion, given these premises - but this result can only be arrived at through the 'testing' process. Johnson-Laird's model, therefore, suggests the following spread of likely conclusions to these premises:

Some A are C
(Some C are A)

No valid conclusion

(here, the dotted line shows the point results below which have required testing - see Johnson-Laird and Steedman 1978, Appendix, 95). The salient quantitative prediction is that the bracketted conclusion will be comparatively rare, due to the 'figural effect'. But now consider the 1II syllogism, *viz.*

Some A are B
Some B are C

which differs only in the propositional form of the first premise. Here the model is

a a
↓
b (b)
↓
c (c)

and its predictions are identical; also, there is no valid conclusion. We have little reason, then, to suspect that the performances of subjects will differ much on 1AI and 1II premises. But they do. According to Johnson-Laird and Steedman's data (*op. cit.* 95) 16 out of 20 get the latter syllogism right (ie. they see that there is no valid conclusion), while 12

out of 20 get the former wrong. This suggests that the 1AI is *much more difficult* than the 1II. Another such anomaly occurs between the 1AO and 1IO syllogisms: here 16 again get the second correct, but only 5 get the first, while 14 get it wrong. Why might this be?

Johnson-Laird suggests that 'erroneous responses may occur if there is a failure to test exhaustively' (64), and in these cases he is clearly right: what happens, if his model is correct, is that in the 1AI and 1AO cases the IR is not tested, and the conclusions given are just those which the model says this will yield, complete with figural effect. But in the 1II and 1IO cases this generally does *not* happen; the model *is* tested; we *do* get the frequent response that there is no valid conclusion. While, in a few cases (3 and 4 respectively), subjects in fact give the wrong answer, and it is then consonant with the IR and the figural effect, the vast majority do not.

Further analysis of Johnson-Laird and Steedman's data shows that conclusions which on Johnson-Laird's theory involve such failure, overwhelmingly occur (a) in the first two figures and (b) when there is an A premise. If both of these conditions are satisfied, 75% of subjects produce a conclusion which appears in the IR, whereas if (b) is not, only 13.5% do so.

Here, there is an interaction between the formal characteristics of syllogisms in general, and the data. It so happens that no syllogism has a valid conclusion which could be read off the IR, unless it has an A premise (see Table 2 in Appendix). All other syllogisms have no valid conclusion, or (if they have EI or IE premises) the conclusion is an O proposition; in either case, reaching the 'correct' conclusion requires testing of the model. Strikingly, the majority (76%) of subjects succeed in doing this. The EI and IE syllogisms are a good deal more difficult than the others, on this showing, especially in the first two figures (which is plausibly explained by the figural effect - see Johnson-Laird and Steedman *op. cit.* 72, for a discussion; interestingly, the incorrect answers here are generally *not* those suggested by the IR.) For syllogisms other than these, and without A premises, 81% of subjects succeed; ie, such syllogisms are *very easy* (*cf. op. cit.* Table 5). However, what the data show is that where there is an A premise in the first two figures, subjects tend to give an IR conclusion *whether it is valid or not;* of the 75% who give such a conclusion, only about half (38%) have produced a valid one. This is what I call the 'A-premise effect' (or just 'A-effect' - see Table 1). It is much more marked in the first two figures. In the second two only 60% of subjects give an IR conclusion, and this derives more from the fourth figure than the third. Oddly, there are *fewer* syllogisms in the first two figures, which yield valid conclusions from the IR, than in the second two.

* * *

What is required is a theory which accounts for these results. There can be no question that Johnson-Laird's theory, as it stands, offers an elegant explanation of the figural

effect; but can it account for these other phenomena? Johnson-Laird and Steedman offer three main kinds of predictions in order to evaluate the theory.

1. Firstly, there is the prediction that the need for testing increases difficulty. However, the figures, already quoted above, are misleading in that the number of problems with correct IRs is small (14 out of 64) and for a large, nonarbitrarily delineable subset of the rest (*viz.* those without A premises and not EI or IE), 81% of responses were correct despite having required testing.

2. There are differences in accuracy predictable 'within the set of problems that are unmodified by logical testing', ie. where the IR yields a correct conclusion, due to the effect of figural bias. Where there is *no* valid conclusion, the existence of bias is supposed to make the problem more difficult because 'the easier it is to form paths, the harder it will be to appreciate that there is no valid conclusion' (84).

3. It is claimed that 'it should be easier to destroy an erroneous initial representation when there are fewer paths to be broken, that is, when premises are particular rather than universal' (84). This appears well borne out in the data. It fits well, of course, with the A-effect, since A-premises are affirmative and universal. The question is whether it suffices to *explain* the A-effect.

It looks as though, if we accept the form of Johnson-Laird's model thus far, *something* must inhibit the testing of the IR when an A premise is present. (This is also suggested by the latency data, as noted above.) What is ponderable, perhaps, is whether the simple consideration of the number of paths in the IR is powerful enough to account for the scale of the differences in the data. Does *one* extra link explain the differences between responses to these problems? We have also the question why the effect of an E premise is much less significant than that of an A, while they are equally universal: the presence of an E premise in a problem does little to raise the number of IR conclusions tendered. Furthermore, this approach ought surely to predict the greatest number of IR conclusions for problems with two A or two E premises, or some combination of these, ie. those where both premises are universal. However, in all the EE cases, and both the AA problems where the IR is incorrect, less than 50% of subjects give IR conclusions; also, the 4EA and 4AE cases are two of the five problems with A premises where less than 50% IR conclusions are given.

Johnson-Laird and Steedman fail to consider these points because they put their main emphasis on the question of accuracy. Indeed, for example, fairly few give the correct response to the AA case where there is no valid conclusion (*viz.* 3AA - 40%, see 84) but also few give the IR conclusion, the difference being accounted for by the presence of two levels of testing before the correct result is reached. This naturally means a greater spread of conclusions, but I cannot see that a reason is thereby generated to expect fewer IR conclusions if the difficulty of IR testing is much affected by the number of paths, which here

is maximal. In any event, Johnson-Laird and Steedman ought to anticipate much greater difficulty in the EE cases than in fact is found; these are more difficult, in general, than the surrounding problems on Table 2, but only marginally.

A possibility one might canvass is that the matter is complicated by the presence of negative links in the model: are they easier to break than positive links? This could explain why there is no 'E-effect' comparable to the A-effect, although it still fails to explain why we don't then have a massive IR response to the 3AA problem, when in fact it's the same as in, eg, the 1EE case. (But here it might be held that *any* explanation for the A-effect is going to have trouble with cases like this.) Evidently, the interactions between particularity and negativity in the premises are rather complex here, and may account for some of the phenomena in the data for problems lacking A premises. However, it seems implausible on the whole to try thus to account for the overall A-effect.

* * *

Let us now offer a speculative solution to this problem. An interesting fact about the IR conclusions predicted by Johnson-Laird's model, is that for any pair of premises, one of which is of the A form, the form of the putative conclusion is that of the *other* premise. So, eg, the IR for a 2AE syllogism suggests an E conclusion, that for a 3IA suggests an I conclusion, and so on. This is consistent with a certain sort of illicit conversion, or misrepresentation, of the A premise. Suppose that 'All A are B' is sometimes represented in such a way as to *identify* As and Bs. (This is suggested by Chapman and Chapman 1959, and explicitly adopted by Erickson, 1974; see Johnson-Laird and Steedman 1978 88,90.) Now consider, eg, the syllogism 2AI

All B are A
Some C are B ;

in this case, if B and A are represented as coextensive, it will be logically correct thereafter to conclude 'Some C are A'. In effect, A can just be *substituted* for B in the I premise; hence, this policy of 'illicit conflation' of the extensions of the terms in the A premise, will invariably yield a result in the mood of the other premise, exactly as do Johnson-Laird's IR conclusions.

As noted earlier, in Johnson-Laird's terms, the A-effect is describable as consisting in some sort of inhibition applying to the 'testing' process of IRs containing A premises. In general, the illicit conflation presented above provides just such an inhibitor. Given the conversion hypothesis, 1AA premises, for instance, will tend to appear simply as

a a
↓ ↓
b b

The 'b' in brackets is omitted; no non-A Bs are recognised. But now this deprives the testing mechanism of the wherewithal for falsification: *any* IR built up using the structure thus arrived at, will yield a logically unfalsifiable conclusion, the nature of which depends on the rest of the structure (ie. the other premise). Whenever there is an A premise, the possibility of testing depends crucially on the presence of 'the "b" in brackets'. So we might suppose that, in many cases, it is simply omitted.

Johnson-Laird and Steedman in fact consider this hypothesis (*op. cit.* 89), but reject it for obscure reasons. They allude to previous work on conversions, saying

> the fact that subjects will accept the converse of an A or O premise as valid provides no direct evidence for a process of conversion, licit or illicit, in syllogistic inference

- which is true; but perhaps the A-effect does provide such evidence. They seem to believe that 'if such a process readily occurred, it would eliminate the figural effect'. However, this is clearly not a consequence of the present idea, and seems to depend on their taking it as part and parcel of a whole system of conversions, dropping optional elements indiscriminately, which then indeed yields results not justified by the data. But if we suggest that it is *only* A premises which are commonly converted on a large scale, this leads only to the A-effect. A major problem with the suggestion as it stands is that it offers no account of *why* A premises are singled out in such a fashion, and fails to explain its occurrence being greatest in the first two figures. This might become a topic of interest if Johnson-Laird's theory were to be established as a strong paradigm in the field of syllogism research. Incidentally, perhaps, if one supposes a side-effect of the conflation to be a collapse of the directional influence of the arrows in the part of the model representing the affected premise, one can derive the 'grammatical role' results described near the end of the last section, at least for A premises, even in the last two figures.

In any event, if the suggestion can be pressed, it serves to question Johnson-Laird's assumption that reasoning proceeds on the grounds of properly-understood premises being used in a completely rational, if resource-limited way. This would potentially cause considerable trouble for Johnson-Laird's integrated account of competence, rationality and processing, which in the chapters now following we shall proceed to examine in detail.

9

Johnson-Laird's theory of Internal Representation

The central point of the structure of Johnson-Laird's theory is that, in his view, it allows us to see the process of reasoning as being semantically driven. There is no 'mental logic', he says, whereby people translate syllogisms and like problems into some kind of internal notation and then operate upon them with rules such as those of the predicate calculus. What happens, rather, is that people understand what the premises mean and use this understanding to generate their implications. The major part of our overall task, then, is to investigate how we can see the quasi-mechanical theory just outlined as a plausible instantiation of this sort of procedure. This chapter will begin that task by examining what is perhaps most fundamental: the interpretation of the diagrams illustrating the 'models'.

9.1 The Syllogistic Models

We have, as noted above, structures such as

```
a  =  b
a  =  b
      (b)
```

(in which particular case, 'All A are B'). It's fairly obvious how Johnson-Laird wants us to envisage these structures as models. We have here a situation in which there are two As, both of which are identical to a B; slightly less naturally, we have a 'possible B' - we reserve judgement on the question whether or not there are non-A Bs. Johnson-Laird gives us an intuitive explanation of this, as follows:

> Suppose you want to draw a conclusion from the premises:
>
> All the artists are beekeepers
> All the beekeepers are chemists
>
> without relying on Euler circles or Venn diagrams. One way in which to

proceed is to employ a group of actors to construct a 'tableau' in which some of them act as artists, some as beekeepers and some as chemists. To represent the first premise, every person acting as an artist is also instructed to play the part of a beekeeper, and, since the first premise is consistent with there being beekeepers who are not artists, that role is assigned to other actors, who are told that it is uncertain whether or not they exist. In short, a tableau of the following sort is set up:

artist = beekeeper
artist = beekeeper
artist = beekeeper
(beekeeper)
(beekeeper)

> There are three actors playing the joint roles, and two actors taking the part of the beekeepers who are not artists - the parentheses designate a directorial device establishing that the latter may or may not exist. Obviously, the number of actors playing the different roles is entirely arbitrary (Johnson-Laird, 1983 94-5).

(Operations like this, somehow straightforwardly 'internalised', are essentially what Johnson-Laird supposes reasoning to consist in.) This suggests the following interpretation. Each line in the model (ie, each actor) corresponds to a single entity in the modelled situation. Hence the *tokens* in the model stand, not for entities, but for *instantiated properties* (or instances of properties) of entities. The '=' signs indicate that the linked tokens stand for properties which are co-instantiated by one real entity.

This interpretation, it has to be said, is not one which everything written by Johnson-Laird encourages us to adopt. Sometimes, he speaks more as if each token is to be regarded as a separate 'actor' in the tableau, and thus as if one has a plurality of actors representing each actual entity. This kind of thing occurs particularly in his earlier accounts of how the mechanisms he envisaged for operating with these models were able to carry out certain of their operations (see, eg: Johnson-Laird, 1975; Johnson-Laird and Steedman, 1978). It is suggested, indeed, by the notion, mentioned above, that we have objects related in certain ways, and in these cases (as it happens) related by *identity*. This, however, entails a curious doctrine about identity and its representation, and one which plays no role elsewhere in Johnson-Laird's theory; hence, I believe that if some serious problems are to be avoided, it is best to cleave to the more recent accounts. We shall see that there are difficulties enough even there, and find independent reasons, anyway, for supposing that relations other than identity require a more complex treatment than Johnson-Laird proposes.

It should be made clear that the interpretation suggested is not one which Johnson-Laird anywhere explicitly lays down. In particular, the notion that each line in the model

tells us about a certain entity is inconsistent with his usage. For example, he draws the model for 'Some A are B' like this:

a = b
(a) (b)

whereas on my account one would expect:

a = b
(a)
(b)

or something similar. This is frequently unproblematic, if we acknowledge Johnson-Laird's tacit convention that tokens refer to the same entity only if explicitly linked. However (and this is something to be taken up later) an added complication is that, often, he seems to treat unlinked items on a single line as being, as it were, 'potentially' linked - somehow more obviously so that those on separate lines - a matter which will be significant in the discussion of the procedures he proposes as operating upon these models.

Johnson-Laird's assertion that the numbers of tokens (actors) in the model (tableau) are 'obviously...entirely arbitrary', will not have escaped notice. This is a crucial part of his story, which is that somehow a model (or tableau), of the sort described, can be regarded as representing in some sense the content of 'All A are B'. The point is that it is a feature of all states of affairs, or situations, in which all A are in fact B, that every A in that situation is identical to some B - no matter how many As there are, nor whether there are non-A Bs. Johnson-Laird holds that we can use an arbitrary number of exemplars of a set to stand for the entire set; thus the three artists in the quotation above represent *all* artists, and the fact that each is linked to a beekeeper indicates that *any* artist one might find is a beekeeper. On the other hand, one *might* find (apart from those whom we know *must be* artists) beekeepers who are *not* artists. Johnson-Laird is also concerned, however, to maintain that a model of the sort in question also represents a particular situation, and that in some circumstances the exact cardinality of the numbers of tokens could be important (eg. if the premise were 'Three artists are beekeepers'). To accommodate these factors, it is necessary to develop an account of how models can be ***used*** in different ways, to model different propositions; I shall therefore now try to explain the basis of Johnson-Laird's views about this, which will inevitably draw in some of his views on language-understanding in general. This will be a fairly free interpretation of Johnson-Laird, who is unclear or inexplicit about many of the points covered. Page references in parentheses indicate places in Johnson-Laird 1983, where support can be found for the statements I make.

9.2 Models and Representation

In the widest sense, Johnson-Laird's theory is based on the idea that organisms (or other systems) whose behaviour is above a certain level of complexity, have to be seen as carrying about with them a model (here in some very loose characterisation) of the environment in which they exist, which model helps them to produce behaviour 'appropriate' to that environment (400*ff*). There are many complex difficulties about this view, many of which are well-known and none of which I propose to discuss here; let us merely note that whatever it is within the organism that constitutes this model, it does so in virtue of its functional role in the production (or control) of behaviour (403). The model can be seen as representational, therefore, just insofar as it has some coherent such role. If, for instance, there is some (neural) structure which operates in such a fashion as to render the behaviour of the organism 'appropriate' to some particular environmental contingency, then (modulo some of the difficulties just explicitly ignored) it can be said to constitute, for the organism, a representation of that contingency - or, if you like, that 'situation'. The level of *detail* in such a representation, is left rather arbitrary.

If the account of such a mechanism involves a particular sort of (representational) structure, it must also specify the arrangements whereby this structure is able to fulfill the necessary controlling functions. This is the familiar 'representation-process' point (*cf.* Anderson, 1978); the representational structure has to be seen as being *'used'* in a certain way. Any biological/psychological structure can be said to represent something *for the organism* (or system) only if it can be shown to play the right sort of role - be used with appropriate sorts of processes - in the functioning of the organism.

Let us, then, consider linguistic cases. We can suppose, as does Johnson-Laird, that such phenomena as language-understanding are related to the evaluation, in some sense, of sentences against a representation of (some relevant part of) the world (156, 245*ff*). We can also suppose that the understanding of any declarative sentence involves the construction of a representation of what it is taken to assert; of a state of affairs in which it's true. A traditional idea, of course, is that to know what a sentence means (ie, I here suppose, to be able to understand it) is to know its truth conditions - in what kinds of circumstances it would be true, and in what kinds it wouldn't - in other words, to have the capacity to distinguish between situations where it is true, and those where it is not. Johnson-Laird takes this capacity to be coupled with the ability to construct models of possible situations in which the sentence is true, and to say of any such models whether they are models of situations in which it actually is true. A model is therefore a structure the use of which, within suitable mechanisms, allows the organism (in this case, the human) to 'pick out' situations of one sort, as compared with those of another; it can thus be said to represent the sort of situation picked out. What is crucial, however, is that the model is not

constrained to be used only in conjunction with one specific type of mechanism, and hence it can be used to represent more than one kind of situation: all that can be said (for reasons which will emerge), is that there is some situation which is simultaneously of every kind that a particular model can be used to represent. Hence the import of a given model in itself is somewhat indeterminate, unless one knows what are the circumstances of its current use. (Familiar remark: the content captured in a model is a function of both the model and the processes which construct and evaluate it.)

Johnson-Laird's story is that when communication takes place, the object is that the hearer should reconstruct in his mind a model identical to that from which the speaker produces the discourse (162-5, 264, 370). This would be best done by a non-deterministic device which always selected the right model from the beginning, but since this is an unreasonable assumption, Johnson-Laird supposes people construct some arbitrary model consistent with the speaker's pronouncements, and later revise it as necessary (assuming for simplicity in the account that the speaker remains consistent in his utterances) (408). He says that when a sentence is heard, one initially produces a propositional representation of it, which serves to disambiguate the surface structure and perform other psycholinguisically necessary functions. This now forms the basis for the model-building processes (244). The representation of truth-conditions is achieved by the resulting model together with these processes, and those which can 'revise and evaluate' the model. Suppose the sentence is 'All the artists are beekeepers', then a proposition of this form is constructed, and the processes produce a model of some situation, no doubt quite unreal, in which there are a number of beekeepers some or all of whom are also artists - and no other artists. (This ignores the question of the 'possible' non-artistic beekeepers, but we'll come back to them later.) There may well be other things in this model, but these are irrelevant details. What is vital is that (a) given this sentence, the processes always construct a model of the sort described, and (b) the processes can revise this model, in the light of information later in the discourse (eg. to the effect that there are exactly five artists in question) so as to make it closer to the one the speaker had in mind. In fact, the processes can produce, originally or by revision, a model representing *any* state of affairs in which the proposition would be true. Notice that what makes one of these models represent that sentence (rather than, say, 'Five artists are beekeepers', or 'Some beekeepers are artists', or 'Some beekeepers are not artists') is the fact that it was constructed, and is liable to revision and interpretation, by just that particular set of processes, operating on the basis of that particular proposition: more will be said later about the implications of this arrangement. In any event, it's easy to see that, even on Johnson-Laird's account, 'Two artists are beekeepers' might well lead to an identical original model, but with different possibilities for revision and interpretation. This is why he says:

> A crucial point about mental models is that the system for constructing and interpreting them must embody the knowledge that the number of entities depicted is irrelevant to any syllogistic inference that is drawn. In the case of numerical or proportional inferences, however, the numbers or proportions will matter. The procedure must accordingly have an independent way of keeping track of which particular proposition is being modelled (98).

* * *

If we want to be a little more formal about all this, what we might say is that, given any model, the processes which constructed and could currently revise it amount to an equivalence relation on the set of all possible situations (states of affairs, whatever). This is not supposed to have anything to do with Barwisean situation semantics, although it may be interpretable in that light (I shall certainly not attempt here to say how situations might be individuated!). The point is that this relation defines a partition of the set into those situations where a certain proposition is true, and those where it is not, and hence, respectively, into those represented by the model under these processes, and those not. Consider, for example, the model

(1)
a = b
a = b
(b)

which has, as noted above, a variety of possible interpretations. What can be said straightaway is that it represents, in the sense just stated, a (any) situation in which there are two As which are Bs, and one B which is not an A (ignoring, again, the brackets). This is, as it were, a rather primitive aspect of its representation. There are two ways in which the story is more complex: the first, which we shall ignore for the time being, is that the model may contains all kinds of default 'world knowledge'-like information about the entities (eg. if they are artists, that they are *male*, *cf. op. cit.* 52-4); the second, here crucial, is that it can be *used* to represent *other classes of situations*. It might perhaps model 'There are two As and three Bs'; it contains, in a sense, too much information, but then it's hard to see how a model fails to represent these as being either linked by some arrangement of identities, or as not being (although not being might be a better default). More clearly in Johnson-Laird's terms, however, there can be procedures which revise it into any otherwise similar model with arbitrary numbers of the same kinds of 'entity-types'. I shall be discussing this matter in more detail shortly, but what it comes to is that the model can be made into any element of the following array (440 - I have added the first column, for consistency with what comes later):

(2)

a = b	a = b	a = b	a = b	
	a = b	a = b	a = b	...
		a = b	a = b	
			a = b	

a = b	a = b	a = b	a = b	
b	a = b	a = b	a = b	...
	b	a = b	a = b	
		b	a = b	
			b	

a = b	a = b	...	...	
b	a = b	...	...	...
b	b	...	...	
...	...	...	...	...
...	...	...	...	

where the elements of this array are all models, just like (1), of particular situations. In this array we have, in effect, a model for every possible situation in which all A are B (for some particular A and B, sets of (instances of properties of) things in the world), and since the revision procedure just mentioned constitutes, if you like, an equivalence relation on this array (considered as a set of models), under which the models are all equivalent, it turns out that (1) is, in an important sense, equivalent to this entire array, so far as its representational ability in this context of use is concerned. Hence, since the models in this array pick out just those situations in which all A are B, we can treat (1) as representing 'All A are B', and the ability to construct (2) as knowledge of its truth conditions. As Han Reichgelt (personal communication) has suggested, it is rather as if one were to treat, say, '4' as representing the integers, given the recursively applicable procedure 'add, or subtract, 1'; in the context of this procedure, '4' is an arbitrary choice.

Every model, taken with a certain set of construction and revision procedures, thus serves to identify a class of possible situations, equivalent just in that they are so identified. Every model, given different procedures, will serve likewise to identify some different class of situations. As suggested above, (1) will pick out situations in which all A are B, or two A are B, or possibly other things. But (1) itself contains a sort of kernel of information which is common to some one situation in each of these classes, their singular intersection; in general it's evident that if (1) represents something under the (2)-generating procedure, then so does anything in (2). Johnson-Laird sometimes speaks of 'representative' models, where the import seems to be models representative of a class of models, as (1) is representative of (2) (165, 190, 264, 439*f*). Evidently, given that they form an equivalence class, any member of (2) will be as representative of the whole set as any other; Johnson-Laird seems to mean, by this usage, just that any member of (2) is, in this

sense, representative of (2) (although, of course, they will differ in what else they are representative of).

As noted above, in determining what a model represents, it is essential to be apprised of the currently operating procedures, or (what is effectively the same thing, in Johnson-Laird's theory) the original proposition being modelled. Without this knowledge, the model might be seen to represent any one of the related, yet disparate, kinds of states of affairs which it could possibly be used to model - and no doubt there can be a very large number of these. Notice, on the other hand, that if one is provided with a proposition, and asked whether a given model could model it, all one need do is look: for, what characterises the equivalence of the states of affairs in the various classes picked out, is really just the truth therein of some proposition, and thus all of these are true in this one model (or the situation which it 'most primitively' models), the nexus of those classes. This notion of 'truth-in-a-model' is just what is captured in the procedures which construct the model from the proposition, and which mediate the inverse operation; it will be used a good deal in what follows.

* * *

As I warned, this has been my own, perhaps idiosyncratic, interpretation of the basis of Johnson-Laird's theory. However, I believe that he has implicitly committed himself at least to most of it. In any event, the assumption that it is correct allows us to make sense of Johnson-Laird's algorithm for syllogisms, in an interesting way; also it provides, to my mind, the only hope of seeing that algorithm as 'essentially semantic'.

In the next chapter, I shall present an analysis of the algorithm, designed to uncover its logico-semantic structure, whereafter we will be in a better position to investigate these semantical claims. In the remainder of the present chapter, partly as helpful preparation for this analysis, I wish to say more about the notion, mentioned briefly above, of a 'representative model', and its implications in the particular case of the algorithm for syllogisms.

9.3 Representative Models

It was mentioned that the notion is founded upon the idea that a model, eg. (1), can be 'revised' into any member of a certain array of models (perhaps better thought of as simply a *set* of models, the arrangement of these being theoretically immaterial), eg. (2) being the array (or set) corresponding to (1). Let me call this array the 'expansion' of the model, and note that in general any model will have some determinate expansion. The object of this apparatus is further to elucidate the way in which Johnson-Laird is able to claim that his models represent the content of a statement, in some helpful sense (in a given context of use).

The expansion, of course, contains models of all possible states of affairs in which the original proposition is true (this much is clear from the above discussion). Hence, it seems not unreasonable to allow that the system, consisting of (1) together with the procedures which expand it into (2), does indeed capture the content of 'All A are B'. We, in that event, accept the following principle:

P: If a model in a particular system is supplemented with processes which allow it to be made equivalent (somehow isomorphic) to any arbitrary one of the possible states of affairs in which a statement is true, then it can be said, in that system, to represent or 'stand for' all of those, and thus capture the truth-conditions and the logical content of that statement.

Unfortunately, Johnson-Laird says that (1) is a *'representative sample'* from (2) (440). This is plainly false. (1) does not occur anywhere within (2); at best one might want to regard it as some sort of disjunction of two of the elements. The problem here is the 'possible' B, discussion of which was deferred from some time ago; not only does it have the consequence just noted, but, of course, it also makes it very difficult to see how a model such as (1) can be regarded as modelling a *particular situation* when, it seems natural to say, it tries rather to sit on the fence between two which are quite distinct. There is, I think, only one way round this: we might suppose that (1) does in fact represent a particular situation, *viz.*

(3)
a = b
a = b
b

and that (3) is itself quite capable of being expanded by some suitable procedures, which allow the omission of unlinked 'b's, into any member of (2). We then suppose that the brackets are there in (1) to 'remind' the procedures that they are allowed to do this. (Presumably also to remind us!) The procedures, that is, are sensitive to the presence of the brackets, which constitute a 'label' on the model of this particular situation that amounts to an instruction to treat it as being one of a certain kind - as expandable in a certain way. (Notice that if the optional B is left out, then there will be insufficient information for the procedures to construct the whole of (2) - *cf.* the discussion at the end of the last chapter). On this interpretation, then, the brackets in (1) have no semantic significance, so far as the specific situation modelled is concerned. Such a view, it has however to be admitted, is not encouraged by much that Johnson-Laird himself says.

If we have accepted P, I contend that the way is open for us to see all of Johnson-Laird's models as being coherently representative of the content of utterances. Perhaps the most doubtful case is that of the representation for 'Some A are not B', the diagram for

which generally appears in its more recent form as

(4)
$$\begin{array}{l} a \\ a \\ \hline (a) = b \\ \qquad\quad b \end{array}$$

where we can see this as, in itself, being a model of a situation in which there are four objects - two As, an AB and a B. The 'barrier' is as irrelevant to this as are the brackets. Clearly it is true of this situation that some of the As are indeed not Bs, although of course all sorts of other things are true as well. The function of the brackets is to allow the expansion of this model into others in which there are no ABs (ie, ones where no A are B), and the function of the barrier, in capturing negation, is to prevent its expansion into any model wherein all A are B (by ensuring that some A, at least, can never be linked to any B). All other possible expansions, though, are permissible, including ones where, say, all B are A, etc., perhaps with the constraint that there have to be some Bs. For instance, the state of affairs (3) would then appear as part of the content of 'Some B are not A'. Hence, it seems reasonable, under P, to say that (4) captures the content of 'Some A are not B', given the appropriate expansion procedures.

We can regard the foregoing as establishing two different kinds of expansion procedure. Suppose we view a model, in the manner sketched above, as consisting of a set of entities, one to each line, of different 'types' (eg. 'a = b', 'b', etc.), then we can say that

E1: The number of entities of particular types can be changed independently and arbitrarily in the range 0 (if they have brackets) or 1, to infinity, and

E2: The types can be changed (entities of other types can be constructed from those present) in certain specified ways (eg. link-making and -breaking).

Evidently E1 suffices to generate the whole expansion of (1), *viz.* (2), but not to generate that of (4). In fact, I think (for reasons which will emerge) that this is an error on Johnson-Laird's part, and that (4) should be replaced by

(5)
$$\begin{array}{l} a \\ \hline (a) = b \\ (b) \end{array}$$

where expansion by E1 will suffice. (Notice here that the apparent type '(a)= b' is just an omissible 'a = b', and that the barrier now does no real work, since its function properly relates only to E2 and the drawing of conclusions. Also, the constraint may yet be desirable that there must be at least one B. The reduction from two to one of the number of 'a's above the barrier is of course immaterial, and is merely introduced for simplicity.)

9.4 Combining Models

I want now to move on to the question of the combinations of syllogistic premises. When one has two premises and combines them, what presumably ought to result is a model which represents, in the sense under discussion, the truth-conditions of the *conjunction* of those premises. What might give rise to difficulties is the fact that frequently one can construct, using Johnson-Laird's recommended methods, up to three different such models. (In terms of the discussion above, this is because one categorical sentence is supposed to 'pick out', by its truth-conditions, one particular kind of situation - those with models in the expansion of its normally related model, in fact - but the conjunction of two can pick out situations different even from those picked out by either of them.) On the view here advanced, this is not a problem. Provided that there exists for each combined model a set of procedures which maps it into the array (analogous to (2)) representing the truth-conditions *of the conjunction*, it is quite possible for each of these models to represent the content of the combined premises. This is guaranteed if P is correct. We might then call the models *coexpansive* under those procedures.

Now, Johnson-Laird's idea is that the reasoner constructs a model to represent the conjunction of the premises of his syllogism, and then is able to 'read off' a conclusion from it. This conclusion is always consistent with the premises, but it may not follow validly from them, so 'testing' has to be done, to see whether other models provide cases where the conclusion is not true. He says that only three (at most - usually less) different models have to be examined. This seems to indicate that examining three models is in some sense equivalent to checking the entire (2)-analogue array for the combined premises, to ensure that it contains no state of affairs in which the conclusion is false. Such would be the case if there were at most three sub-arrays, as it were, each equivalent to the expansion of one combined model, in each of which the question of what was true in all of its components could be decided by examination of a single model of which it was an expansion.

The suggestion is, then, that there should be some natural procedure which can expand combined models into sub-arrays, but also some other which expands these models *into each other*, in the appropriate sorts of ways. It would be initially plausible to suggest E1 and E2 above, respectively, for these roles.

The idea is that the models are interderivable by expansion under procedure E2 above. What then needs to be shown is that the models thus produced suffice, when expanded under E1, to generate the whole (2)-analogue array for the conjunction. (Alternatively, we could see the effect as arising from *construction* rules governing how the two representations are combined: there should be at most three ways of doing this which do not yield models that expand equivalently under E1.) Thus if we say that a *model-class*

consists of (a maximal set of) models interderivable by (ie, coexpansive under) E1, it follows that the claim here is:

C: (i) The (2)-analogue arrays representing the contents of conjunctions of syllogistic premises, contain at most three distinct model-classes, and (ii) these correspond to Johnson-Laird's combined models.

It will now be seen that the point of my amendment to (4) was just to ensure that each premise's single content contains only one model-class. One could then say that E1 defines an equivalence relation on the array, partitioning it into *n* (not more than 3) disjoint sub-arrays (E1-interderivability equivalence-classes; model-classes).

In the next chapter, I shall be arguing, among other things, for the truth of the first part of C; the second part seems more dubious, but will be considered also. For the present, I shall summarise the latest results of the discussion.

9.5 Summary of the Theory

If the argument is correct, it follows that a set of exemplars of the model-classes for a syllogism will constitute up to three models of different specific states of affairs, the summed expansions under E1 of which yield the content of the conjunction of the premises, as does the expansion of any one of them under E1 and E2 combined. Imagining the models of each model-class arranged in vertical columns, E1 provides a way of moving vertically up and down these, while E2 allows one to move horizontally between columns. Thus, given both, one can move from any model to any other; the two together form an equivalence relation equating all the models in the (2)-analogue array. (Han Reichgelt suggested to me that this might seem clearer expressed as follows: suppose, where *m* is an arbitrary model, that E1(*m*) yields sub-array A; E2(*m*) yields *m′*; E1(*m′*) is A′: then the union of all A′, A″,..., thus obtainable, is equivalent to the whole array.) In one or other of these states of affairs, it will be found that any syllogistic proposition consistent with the premises is true (given a usual sort of interpretation function); moreover, if *and only if* any proposition is true in *all the E1-expansions of* all of them, ('true in each model-class', one might say) then it's true in every situation in the (2)-analogue of the conjunction, and hence is a logical consequence of the premises. Clearly, there will be propositions true in some, but not all, of these models (where there are more that one). Thus, in other words, models of wholly different situations can be *used* (by the system) with the right procedures, to represent the same content; and, of course, the same situation can be used to represent different contents - *cf.* these, for 'Some A are B', and 'Some A are not B', respectively:

(6)

a	=	b		a
(a)				———
(b)				(a) = b
				(b)

Knowledge of the procedures involved is essential, and is supposed to be to some extent conveyed by the brackets and barriers.

This is how, in principle, Johnson-Laird is able to predict the drawing of different conclusions from different models, derived from the same premises, all of which might be seen to represent the same truth-conditions. The drawing of a conclusion (in his sense) is just the uttering of some particular statement which is true of the model it is based on: it is an attempt (partially) to describe that model, or a situation it is taken to represent. Clearly, such a statement may be true of no other model-class, and hence not implied by the premises.

* * *

We now have some idea of the sort of background against which Johnson-Laird's algorithm is to be examined. It is important to realise, however, that little of this is anywhere made explicit by Johnson-Laird himself, which is why I have had to labour some of it, perhaps at seemingly excessive length. And I am afraid to say that we have by no means finished with it yet. There remains the problematic matter of accounting for which conclusions are actually drawn from which models. This evidently has to depend somehow on the distribution of tokens, brackets and barriers. Moreover, we need some explanation of how the 'expansion' procedures, referred to above, can be seen to be relevant to Johnson-Laird's algorithm for syllogism-solving, which appears to involve no such apparatus. There's a clue to the connection of these, in the observation that one's conclusion drawn from a particular model has to be true in all the models of the appropriate model class. In the following chapter, we shall try to arrive at an explanation of how Johnson-Laird's algorithm emerges from the present theory as an *effective procedure* for deriving valid conclusions to syllogistic premises.

10

An Approach to the Basis of Johnson-Laird's Algorithm

In the last chapter, we derived a theory about models, according to which each can be regarded as representative of a certain subclass of all the models in which the premises are jointly true, and how it might then be shown that each successive model represents a different, disjoint such subclass, until the entire class has been covered by at most three such models. A claim 'C' was advanced, stating the main point of this theory, and also that the models involved correspond to those used by Johnson-Laird. We are now concerned mainly with the defence of that claim.

A difficulty arises with matching the account given there to Johnson-Laird's models, because the combined models needed for the suggested procedure 'E1', are not like his. Consider the case of the 1II syllogism

Some A are B
Some B are C

Here we need, for E1 to generate one of the two model-classes, the following model:

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

whereas all Johnson-Laird has is

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

which is useless to E1. One might attempt to deal with this by proposing some additional expansion procedure which is capable of adding the three missing entity-types, but a better solution seems to me to be to direct attention to the fact that the *interpretation* of models is so crucial in determining what they represent. In syllogisms, this is essentially carried out by the set of conclusion-drawing procedures. We can define a model-class as that subset of models of the combined premises from which just the same conclusion(s) can be drawn.

10.1 Model-classes and Conclusions

In the story sketched above, the intuitively essential aspect of a model-class is that it contains only those models in which a particular proposition (the putative conclusion) is true simultaneously with the premises. But not necessarily all of these: a valid, multi-model syllogism must have a conclusion true in all models and, correlatively, a given model can support more than one putative conclusion. In the context of Johnson-Laird's theory, we think typically of there being two conclusions, one read in the A-C direction, and one in the C-A. In order to see that either conclusion is *incorrect* (fails to complete a *valid* syllogism) it is sufficient to show the existence of a model in which the premises are true but it is not; for it to be correct there must be no such model: however, the production of a new model-class, in itself, *guarantees* neither of these for both conclusions. On the other hand, a distinct new model-class must surely be such that at least *one* of the putative conclusions is not true in all of its models. This being the case, if we assume that the conclusion-drawing procedures will 'read off' a certain proposition only from a model in which it is true, the way is open for the definition of a model-class as

D: a model-class is that set of models in which both premises are true (ie, models constructable from the models of those premises), and which are indistinguishable from one another by the set of conclusion-drawing procedures (are equivalent in those procedures' eyes).

(This is a slightly more formal version of the implicit definition of a model-class suggested by Johnson-Laird and Bara, 1983, 49.)

The essential components now required, therefore, are

(a) some procedure(s) for constructing combined models of the premises,

(b) some procedure(s) for reading off conclusions from these models, and

(c) some procedure(s) for 'transforming' the models in ways which preserve the consistency of the models with the premises, but destroy their consistency with some conclusion which was or might have been offered by (b) - *viz*, at least one of the putative conclusions.

These components, of course, are just those which Johnson-Laird adumbrates. Clearly, the *application* of (b), if (c) does not essentially appeal to its actual output, is an optional action at all stages but the last.

What has now to be shown, then, is that the procedures of the kinds (a), (b), and (c), offered by Johnson-Laird, are such as to suffice for the tasks allotted to them. It is not enough, remember, for the procedures merely to produce, even in all cases, the correct conclusions, if they do this in any sense 'by accident': they have to be seen to do it *in the right sort of way*. Since it is obvious that if a putative conclusion is false in any new model produced, then it must be rejected as a possibility, what really has to be shown is that model-classes as now defined can actually do the job of identifying *valid* conclusions, in a principled way.

If we consider the last chapter's explanation in terms of 'expansion procedures', which seemed entirely coherent, it seems reasonable to proceed by showing that the model-classes defined in this new way coincide with those defined in that way. This obliges us to look at the conclusion-drawing procedures in some detail; they are defined thus (see Johnson-Laird and Bara, 1983, 46-7):

> The theory distinguishes four possible relations in a mental model between a token of one end term and a token of the other end term:-
> 1. There are positive links between them . . .
> 2. They are completely separated by one or more 'impenetrable' negative barriers . . .
> 3. A negative barrier separating the end terms may be 'penetrable' in that it has members of the same class (either an end term or the middle term) on both sides of it [otherwise it is 'impenetrable'] . . .
> 4. The end terms may be in an indeterminate relation, i.e. they are neither linked positively nor separated by a negative barrier.
>
> The principles underlying the formulation of conclusions are straightforward.
> [1′(a)] If there is a positive link from each token of A to a token of C, the conclusion has the form:
>
> All the A are C.
>
> [1′(b)] Otherwise, if there is at least one positive link from a token of A to a token of C, the conclusion has the form:
>
> Some of the A are C.
>
> [2′] If all the tokens of A are separated by at least one impenetrable negative barrier from the tokens of C, the conclusion has the form:
>
> None of the A are C.
>
> [3′] If the negative barrier is penetrable (as defined above), then the conclusion has the form:

Some of the A are not C.

> [4′] Finally, if there are only indeterminate relations between the end tokens, then there is no conclusion that can be drawn interrelating them. These principles yield . . . the maximally informative conclusions consistent with the models.

These definitions are clearly related to what one might want to see as the 'natural' (certainly, traditional) way of interpreting syllogisms in terms of intersecting sets - and of course this is intentional. The 'entity-types' mentioned in 1 are clearly indicative of an intersection between all three sets (and therefore between the two 'end' ones), 2 relates to a complete lack of any intersection between the extensions of the end terms, while 3 shows that one end set intersects both with the other set and with its complement, and in 4 there is no definite such information. It hence follows that the conclusions drawn as described on these bases are well-motivated. Let us then consider whether we have here sufficient for a proof that the procedure as a whole is an effective one for solving the domain of syllogistic problems in which Johnson-Laird is interested.

10.2 Johnson-Laird's Domain of Problems

In order satisfactorily to address this question, we should first satisfy ourselves as to what Johnson-Laird's domain of syllogistic problems actually is. This is something which is commonly taken entirely for granted by psychologists, who cite some such basic work as Cohen and Nagel (1972) as providing a suitable account of the sort of problems they are dealing with, and say no more about it. In the context of demonstrating effectiveness for a solution procedure, this is clearly inadequate.

In fact, Johnson-Laird is much clearer than most in specifying his problems. As a rough approximation, one can say that he is interested in the area of Aristotelian syllogisms, restricted to those which (a) are categorical, (b) have exactly two premises, (c) have both premises and the conclusion in the form of strictly AEIO propositions, (d) involve exactly three terms, one of which occurs twice in the premises, but not at all in the conclusion. That is, he says nothing about any inferences which might involve hypotheticals, or disjunctions, etc.; he does not attempt to treat the cases of inference from only one premise (*immediate inference*), or from three or more (*sorites*); he says nothing about inferences which involve quantifiers other than the simple universal and existential, in combination with negation, nor about traditional contrapositives, obverses, etc.; and certain quite possible inferences are ruled out of his system as simply ill-formed and beyond its scope. (For example, inferences such as

All B are A
All B are C

All B are A and C

and

All A are B
No B are not C

All A are C

are not accounted for.) Therefore, his theory is restricted to traditional syllogisms like those of the Scholastics, a domain of precisely sixty-four kinds of problem, with no obvious and simple extensions. We noted at the outset that he does not treat these in the same way as the Scholastics - his choice of figures, for instance, is different from theirs - and of course there are four possible conclusions to each syllogism, giving a total of 256 configurations. A useful practice of Johnson-Laird's is to lay out these facts in some detail in each of his expositions of his theory.

One can point out facts such as, for instance, that the second of the syllogisms just given is simply equivalent to one of Johnson-Laird's, under a straightforward transformation of the second premise. However, insofar as the theory is a *psychological* one, and says nothing about how such a transformation might be realised, the omission is more significant than it would be in a purely logical theory. Johnson-Laird says elsewhere than in his detailed discussions of syllogisms, that he expects his general theory of mental-model usage to be extendible in like fashion to several of the other cases we have mentioned - but he makes no concerted or detailed effort in this direction. In particular, the form of the mechanism under consideration clearly makes no attempt of any kind to extend at all, and we shall see that the idea of extending it runs into serious difficulties.

A formal characterisation of Johnson-Laird's problem domain exists as a subset of the formalisation of the Aristotelian syllogistic as a system of proof-sequences, provided by Smiley (1973). Smiley's system is rather general, in that it accommodates any number of premises, but of course two will fall out as a special case. Smiley accounts for the particular set of syllogisms which arise in Aristotle's (and Johnson-Laird's) system, by introducing the notion of a *chain* of formulae, as follows.

> Let Vab stand indifferently for any of the wffs Aab, Aba, Eab, Eba, Iab, Iba, Oab, Oba. Then by a *chain* of wffs I intend primarily a sequence of the form $\langle Vc_1c_2, Vc_2c_3, \ldots, Vc_{n-1}c_n \rangle$ . . . Thus to say that a set of wffs is a chain linking a and b is to mean that either its members can be arranged in a sequence of the kind described, with a as c_1 and b as c_n, or else that it is empty and a = b . . . eg, the premises in Cesare form a chain Aab, Ecb linking the

terms a and c of the conclusion Eac. (Smiley, 1973, 144.)

With this in hand, Smiley is able to proceed to the following crucial definition:

> X and Q belong to an *Aristotelian mood* if they can be derived by simultaneous substitution of terms from X_1 and Q_1 such that (i) X_1 is a non-empty chain of wffs linking the terms of Q_1, and (ii) no term occurs more than twice in X_1, Q_1. If in addition Q_1 is a logical consequence of X_1, the mood is *valid*. (*Loc. cit.*)

Hence we have a more principled way of restricting the domain of problems to just those in which Johnson-Laird in fact interests himself. It is interesting, perhaps, that Smiley is forced by his fidelity to Aristotle to mirror many of Johnson-Laird's rather similar precepts; eg, the omission of the 'syllogisms of strict implication' (Smiley, *op. cit.* 139), which is the basis of the 'chain' theory - cf. Johnson-Laird's strictures about 'uninformative' conclusions (eg. 1983 34*ff.*). It thus appears that we can rigorously specify the logical nature of the domain of problems which Johnson-Laird's theory is supposed to address (without here asking whether the restrictions on it make any psychological sense).

10.3 Syllogisms and Decidability

It will be useful to employ a result stated on a different occasion by Smiley (1962). Here, he interprets syllogistic logic in terms of a many-sorted predicate logic, and is thus enabled to show that it is decidable in the 7-element domain, as follows.

> A wff containing letters (predicates or variables) of not more than *n* sorts is a theorem iff it is true under every interpretation in terms of the non-empty subclasses of a domain of 2^n-1 individuals. The justification of this runs parallel to the justification of the corresponding decision procedure for the ordinary singulary predicate calculus [Church, 1956, **466]: in any interpretation we can class together those individuals which belong to the same selection of the *n* domains involved, and since every individual must belong to at least one domain there are at most 2^n-1 classes to be formed in this way. Then we show that the individuals in each class can be lumped together without affecting the truth-values which wff. receive under the interpretation.

(Syllogisms, of course, have three terms, and so the relevant domain contains 7 individuals.) This result shows that an effective procedure for solving syllogisms is at least something it makes sense to look for.

Smiley's method can be clearly related to the story in the last chapter. The 'individuals which belong to the same selection' are just entities of the same 'type', as there defined; hence there are at most 7 entity-types in a syllogistic model. This is evident enough if one considers that an entity of a given type is just one which is contained within the extensions of a certain set of predicates, or has certain properties. Suppose we reformulate Johnson-Laird's models as simply sets of entity-types, which for present purposes is

all they are: they can then be generated by selective deletion from the set of seven possible types. A perspicuous notation for types might be such as to use, eg, '<A,B,C>', for an entity which has all three attributes at issue. But for compatibility with what has gone before, I shall persist in writing this as 'a = b = c'. We then have it that the set of seven possible entity-types is

(1) {a = b = c, a = b, a = c, b = c, a, b, c}.

A pair of syllogistic premises will rule out zero or more of these. 'No A are B', for instance, rules out the first two; the 1II premise-pair rules out none at all; 1AA reduces them to

(2) {a = b = c, b = c, c}.

This resultant, reduced set (call it the *type-set* for the syllogism) will generally contain some subsets (including itself) which are possible *models* of the premises. By a model, generally speaking, I mean simply any set of types (ie, any subset of (1)), but a model *for a premise-pair* clearly has to be restricted to being a subset of the type-set for those premises, with the additional constraint that each of the three attributes must be included in at least one entity-type (since the sets are assumed non-empty). Thus, something like

(3) {a, b, c}

is not a model for any syllogism containing an I premise, but is a potential candidate for some combination of E and O premises. The collection of models for a given premise-pair, I shall call the *model-set* for those premises.

The model-set for the 1II case is clearly very large indeed, but in the more manageable 1AA case, for instance, we have (other than the type-set itself)

(4) (i) {a = b = c}
(ii) {a = b = c, b = c}
(iii) {a = b = c, c}

but not just {b = c}, or {c}, of course.

Let us now adopt the following convention. If a model in the model set contains a type the omission of which leaves that set equivalent to another member of the model set, that type can be 'marked' (eg. by enclosing it in brackets), and the smaller model thereafter ignored. (Such marking is not obligatory wherever it is possible, however.) It must therefore be possible to omit any or all of the marked types in a model, without thereby producing a non-model, but none of these actions shall be obligatory. A model with marked types can thus be regarded as an abbreviation for the set of models consisting of itself and those of its subsets obtained by restoring the possible omissions.

In the 1AA case, then, we are left with the model set written as

(5) {a = b = c, (b = c), (c)}

which looks very familiar, and indeed it is natural now to suggest that Johnson-Laird's models are simply equivalent to such subsets of the model set for a given pair of premises, with appropriate application of this abbreviating convention. Notice that these models could be 'expanded', by use of the procedure E1, described in the last chapter, without prejudice to their representational capacity, since they would still contain only the same entity-types; it would thus be reasonable to regard them (commensurately with the terminology for their components) as *model-types*. However, there are certain important points yet to be covered.

10.4 Conclusion-drawing Reconsidered

The contention is that the model-set for any syllogism can be partitioned into at most three subsets, the members of which are equivalent in that they are treated identically by the conclusion-drawing procedures. (They then count as model-classes, in the terms of the above discussion - see definition D.) To the end of showing this, I reformulate those procedures (without, I think, doing violence to their original statement above), as follows, understanding that the entity-type 'a = c' is said to 'contain both *A* and *C*'.

1(a)˝ If all the entity-types in a model, which contain *A*, contain also *C*, then the conclusion is 'All A are C'.

1(b)˝ Otherwise, if at least one type containing *A* contains also *C* then the conclusion is 'Some A are C'. (However, it will be noticed that here we may allow the reading off of this conclusion even in cases where both premises were negative; hence it is necessary to add the condition: *iff there is no entity-type* 'a' *in the model.* This captures, in effect, a major function of Johnson-Laird's 'barrier'.)

2˝ If *A* appears as contained in 'a' or 'a = b', then the conclusion is 'Some A are not C', unless the antecedent of 3˝ holds.

3˝ If *A* appears *only* in 'a' or 'a = b', then the conclusion is 'No A are C'.

4˝ If none of these obtain, no conclusion can be read off.

It will be noticed that the 'marking' of types is irrelevant in all of these definitions. I hold it to be obvious on inspection that they are correct, in the sense that the conclusion will be true of the specified sort of model considered as a description of the relevant aspects of some 'situation' (see previous chapter). Obviously, too, one can rewrite these procedures to give appropriate conclusions going, as it were, the other way, eg. 'All C are A'.

We can now look at the example of the 1EA syllogism. Here the type-set is

(6) {a = c, b = c, a, c}

which generates the following model-set (types re-ordered for clarity)

(7) (i) {a, b = c, (c)}
(ii) {a, a = c, b = c, (c)}
(iii) {a = c, b = c, (c)}.

Given the marking convention, it might seem, all these could have been expressed simply as

(8) {(a), (a = c), b = c, (c)},

but it must be borne in mind that we have to have at least one of the first two types in every model (at least one type containing *A*). That is, we have one or the other, or both, which might be thought to explain the presence of just three models in (7). However, we could quite consistently reduce these three to two, and in one of two different ways, *viz*:

(9) (i) { {a, (a = c), b = c, (c)}, {a = c, b = c, (c)} }
(ii) { {(a), a = c, b = c, (c)}, {a, b = c, (c)} },

but the three in (7) are chosen because they are the ones discriminated by the conclusion-drawing procedures. We have 'No A are C', 'Some A are not C', and 'All A are C' read off, respectively. Reading 'from C to A', we have 'Some C are not A' from all three, and hence it is the correct, valid conclusion. There are therefore exactly three definite model-classes for the 1EA syllogism, the canvassing of which has in some sense involved something equivalent to the canvassing of all possible combinations of the relevant entity-types.

In the 1II case, our marking convention allows the model-set to be written as

(10) { {a = b = c, (a = b), (b = c), (a = c), (a), (b), (c)},
{a = b, b = c, (a = c), (a), (b), (c)} },

which indicates that the members of the model-set must *all* contain as subsets either {a = b = c} or {a = b, b = c}, or both. These two members given constitute the two distinct model-classes appropriate for this syllogism, and allow 'Some A are B' and no valid conclusion, respectively, to be read off. But now consider this relatively impoverished-looking model-set for 1II:

(11) (i) {(a = b = c, (a), (b), (c)}
(ii) {a = b, a = c, (a), (b), (c)}.

Here we have what Johnson-Laird actually produces, and the suggestion is that it is no less adequate than the version in (10). Recall that a model-class is defined, by D, as an equivalence-class, where, within a given syllogism, the equivalence relation amounts to

xRy: y yields the same conclusions as x under procedures 1″-4″. It seems clear enough that in this case the two versions of the 1II model-set contain exemplars of the same two equivalence-classes, the same two model-classes. It matters nought that (10) cannot easily be *generated* from (11); presumably, any model of a model-class will serve in these circumstances as well as any other, from the logical point of view. (From the psychological point of view, of course, it all depends on the details of your theory.)

10.5 The Theory in Relation to Johnson-Laird's

If we accept the notion that each line of a Johnson-Laird model stands for a separate entity of a given type (and we have considered in the last chapter the extent to which this is a justifiable interpretation of Johnson-Laird's view), then we can see that those models are (apart from some factors which will be mentioned shortly) essentially equivalent to members of the model-set, and thus are exemplary model-types from the particular model-classes for the relevant premises. (There is in fact a striking similarity between the notation I have used above, and the constructions produced by Johnson-Laird's unpublished LISP-80 computer program.) My use of the 'marking' convention makes this clearer, since it corresponds to aspects of Johnson-Laird's notation, but it has no logically central function in the argument. Since it seems clear on Smiley's evidence that the machinery introduced above provides a decision-procedure for a certain class of syllogistic problems, and since I think that Johnson-Laird's machinery is in principle much the same, I think his claim to possession of such a decision procedure is to that extent vindicated. It seems clear, moreover, that the relevant class of problems is just that delineated earlier in this chapter. Notice that the procedures for transforming models have played no part in all this: their role is purely psychological, and indeed Johnson-Laird's are merely heuristic, drawn from a wide range of possible alternatives.

Regarding the 'representationality' of these 'models', they can be regarded (in the fashion discussed above) as models of actual-world 'states of affairs', or 'situations', if it is accepted that a set of entity-types can be held to represent any of the class of situations containing those types, such that the given types specify *all* the combinations of the (three) terms in these situations. We have here yet another appeal to the notion of an equivalence-class, and presumably everything about the members of these classes (ie, the situations) which is not germane to the equivalence-relation, is simply irrelevant, and need not be specified. Armed with a set of types, and a suitable definition of situation-equivalence based on these, one has all that is necessary and sufficient to identify any situation in which the premises of a syllogism, say, are true; and each of the multiple models for a syllogism identifies a subclass of these situations. Since this is all laid down in terms of what types there are, with no reference to the actual number of these in any given situation, we automatically have the benefits obtained by the 'expansion' procedures

of the last chapter, only in a somewhat more abstract guise.

10.6 Rationalised Models

At this point we can make several observations. In the first place, Johnson-Laird's actual expositions of the theory (principally Johnson-Laird, 1983; Johnson-Laird and Bara, 1983) tend to incorporate various *ad hoc* adjustments to facilitate computer-implementation, data fitting, and similar non-logically-motivated aspects of the subject. Hence, for example, the models in (7) are altered, as are the conclusion-drawing procedures, so that we get as 'A-C' conclusions: (i) No A are C, (ii) Some A are not C, (iii) No Valid Conclusion - thus avoiding having to incorporate an explicit procedure which notes that 'All A are C' is incompatible with the other two. I have also had to play a similar game in my definition 1(b)´´ above, which suppresses the reading-off of 'Some A *are* C' from (7)(ii). Johnson-Laird's intuitive idea is that people, imagining models of this kind, think to themselves: 'Ah, some A *might be* C, so it can't be that *no* A are C after all. But, still *some* A *are not* C' - until they come to the third kind of situation in which even this is false. But they don't *read off as a conclusion* 'Some A are C', at any point. Why not? Presumably, because they see that this can't possibly be true in *all* the models, in particular the first. (On the other hand, I do not deny people the ability to see that 'No A are C' is false in a model without explicitly 'noticing', or 'realising', that 'Some A are C' therefore must be true. It must though, nonetheless, given the existential-import assumption on which these models are based.)

Consider Johnson-Laird's actual diagrams here:

(12)

1.
```
a
a
-------
b = c
b = c
  (c)
```

2.
```
a
a       (c)
-------
b = c
```

3.
```
a       (c)
a       (c)
-------
b = c
b = c
```

(see Johnson-Laird and Bara, 1983, Table 9). It's hard to see these strictly in terms of my

discussion: they all contain just the same entity-types, and the first two even have the same numbers of them (hence, on my account, representing the same primitive states of affairs). Significant is the point mentioned some time ago, that tokens on the same line tend to be seen as in some sense 'potentially' linked; this occurs here because they are on the same side of the barrier, and hence are not debarred from being linked. This is obscure, however, and I propose that (12) be completely overhauled and rationalised, as:

(13)

1.
a
a
---
b = c
b = c
(c)

2.
a
a = c
---
b = c
b = c
(c)

3.
a = c
a = c
---
b = c
b = c
(c)

These are rather obviously equivalent to the models in (7), modulo the barriers. (*Cf.* Johnson-Laird 1983, 96-7, where the models for 3EI are much as I have just prescribed, albeit, so far as one can tell. more or less accidentally. Johnson-Laird commonly fiddles with the details of his machinery in no very principled fashion until the predicted conclusions come out right.)

The discussion about (8) raises another point noted by Johnson-Laird, which is that if the conclusion-drawing procedures are specified differently, the number of models for each multi-model syllogism can be reduced by one (see Johnson-Laird and Bara, 1983, 62-3), or indeed otherwise adjusted. The way in which he does this is semantically horrifying (and his explanation of how difficulty-predictions still stand seems to me to be, psychologically, equally so), but it does go to show that the actual distinctions introduced above are often fairly arbitrary, and geared to the subject at hand. This, of course, does not diminish the coherence of the procedure, from the logical aspect, but suggests that the semantic story does little to constrain the predictive output of the theory as a whole.

10.7 Conclusions

On balance, then, I am urging that Johnson-Laird's theory presents a coherent algorithm embodying an effective decision procedure for his espoused set of problems. His expositions, and concessions to convenience, tend to marr this, however. He unwisely makes the numbers and positions of tokens (especially bracketted tokens) in his models, into a crucial consideration. This militates against his claims about arbitrariness, and threatens to undermine his semantical story; further, his assignment of numbers of models to premises comes to appear largely arbitrary. But I hope that my efforts at analysis and reconstruction have shown these drawbacks to be essentially surface phenomena; eliminable warts. In the next chapter we must pause, before examining Johnson-Laird's claims to have avoided using 'mental logic', to look at his claims about the detailed structure of his models, beyond their relatively abstract properties we have looked at here.

11

Comparison of Johnson-Laird's Theory with Others

In this, somewhat shorter chapter, we look at how Johnson-Laird's theory compares with others that have preceded it. All these preceding theories are of course known to Johnson-Laird, and in a number of places he argues for the superiority of his own theory. We shall accordingly examine these arguments, and in so doing find them often rather less than compelling. It will be claimed, in particular, that Johnson-Laird's theory, in so far as it is clear, is in fact essentially equivalent to some of these other theories, and that where it may significantly differ it is hard to assess the extent or importance of the differences. This result is of importance for assessing claims about the 'psychological reality' of these models, or in general any claims based on their detailed structure. In this chapter, we assume familiarity with the material presented in the last two chapters.

11.1 Johnson-Laird's Critiques of his Predecessors

(a) Johnson-Laird and Steedman (1978 *op. cit.*) contains a long section devoted to criticising previous theories; those based on 'atmosphere' or 'illicit conversions' are rejected (for the most part justifiably) as insufficiently, even where at all, explanatory, and the notions of Erickson (1974) are retained as the sole serious contender. Erickson's theory depends on the idea that reasoners construct 'Euler circle' representations, and fall into error through missing the complexities of these. People are assumed to use only certain of the (often) very many representations necessary for arriving at the correct conclusion, and assessing various probabilities in their selections can in principle account for most sorts of data. Considered as a cognitive model, this is certainly incomplete and in some respects unsatisfactory; Johnson-Laird criticises it for being unable to reproduce the proper range of 'no valid conclusion' responses observed, for having no 'heuristic' wherewith to *predict* which conclusions should be most probable, and for being unable to cope with the figural effect at all. In this last regard, he has however to concede that if

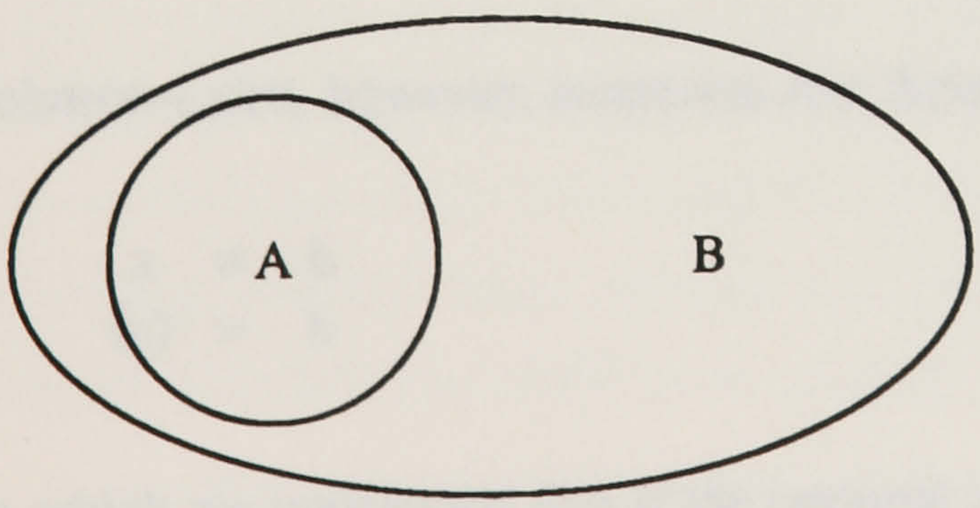
augmented with assumptions about 'the importance of the grammatical subject' in syllogistic premises and conclusions, or about the detailed operation of 'working memory' (91), the theory could deal with the problem. Since Johnson-Laird has also to agree that a suitable heuristic for drawing conclusions could easily be devised (he actually suggests a simple one), he has to admit, as we shall shortly see, that apparently his 'analogical theory' could equally well be based on Euler circles.

(b) Johnson-Laird and Bara (1983 *op. cit.*) contains an even larger section of invective directed at substantially the same competitors (with the addition of Braine and O'Brien, 1983). Atmosphere and conversion being given even shorter shrift, the offensive returns to Erickson, and to other versions of the same theory (particularly Sternberg's (Guyote and Sternberg, 1978), which is attacked in some depth in Johnson-Laird, 1983 *op. cit.*). The criticisms above are reiterated, but with an additional objection to the impeachment of subjects' 'rationality'. Johnson-Laird holds that if no subject is allowed by the theory to arrive at the correct conclusion to any syllogism 'for the right reasons' (ie, in this case, having considered *all* the relevant Euler-circle combinations), then humans must be essentially irrational, and that this is contrary to manifest fact. Such an argument turns on highly controversial matters, as we have seen in Part I; this aside, however, Johnson-Laird's case in the earlier paper that his theory is intrinsically superior in its mode of representation is severely weakened by its being forced, in its latest guise, to rely on just those assumptions about working memory which, it was previously said, would allow Erickson's theory to explain the figural effect. In these circumstances, it is worth exploring a little the possibility that Johnson-Laird's theory could be fairly easily recast so as to employ Euler-type circles rather than individual tokens.

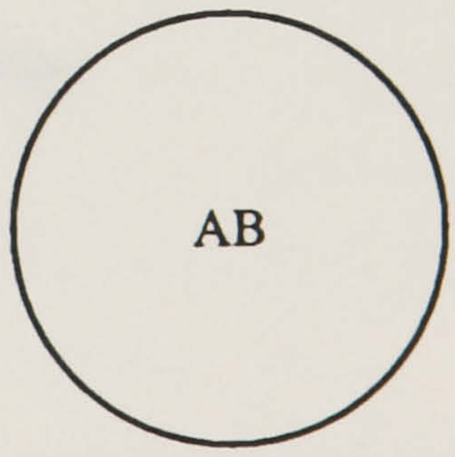
11.2 Euler Circles

In the Euler circle treatment, 'All A are B' has to have *two* different models, *viz.*

I



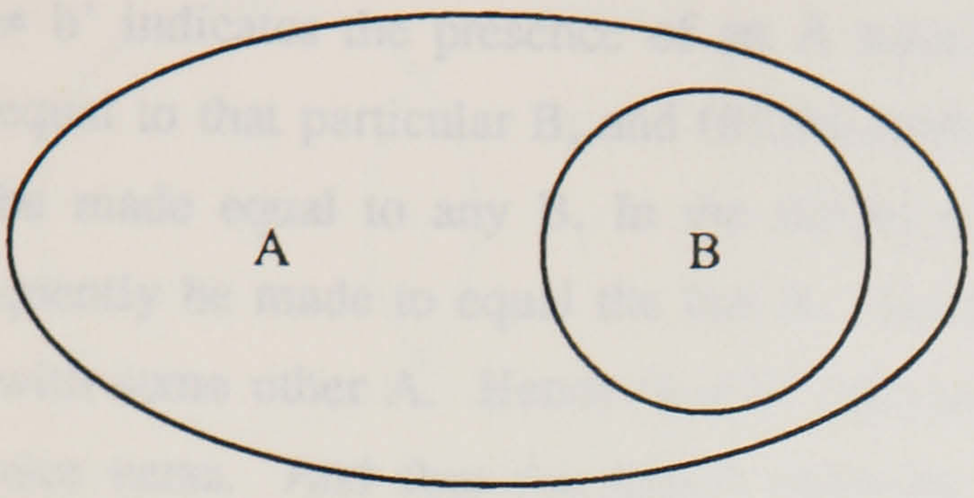
II



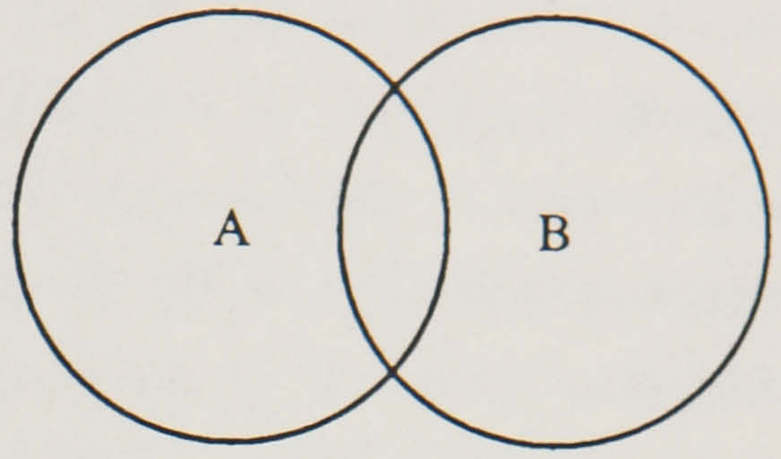
where I is the case when the 'b in brackets' is present, and II is that when it's absent. (It's assumed in these representations, but not in Venn diagrams, that all areas are non-empty.)

'Some A are B' needs both of these, plus these:

III



IV

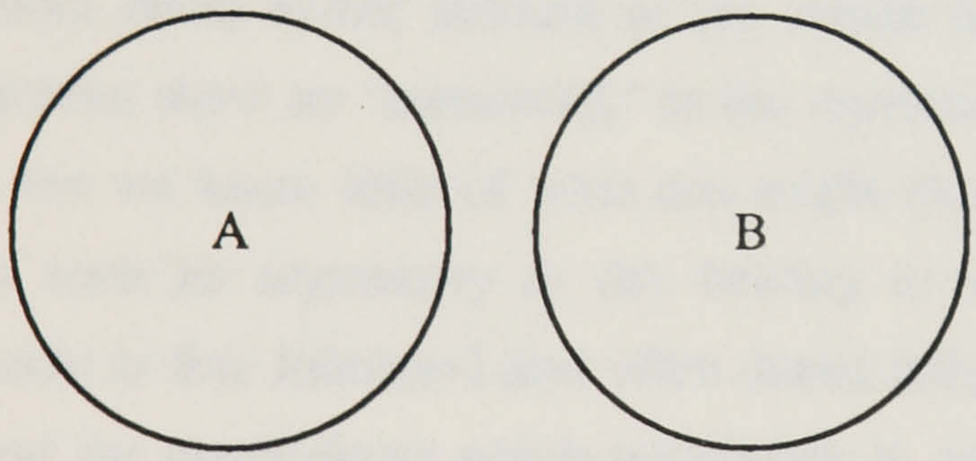


but Johnson-Laird claims to combine all four of these in his representation:

a	=	b
a	=	b
(a)		(b)

(ie, the cases when one or neither, or the other or both, of the optional elements are present). To deal with 'Some A are not B', the Euler circle system needs to consider three possibilities: III and IV above, and also:

V



Johnson-Laird, however, maintains that these are covered by

a	≠	b
(a)	=	b

in which we understand that if the optional A is omitted, the linked B also drops out.

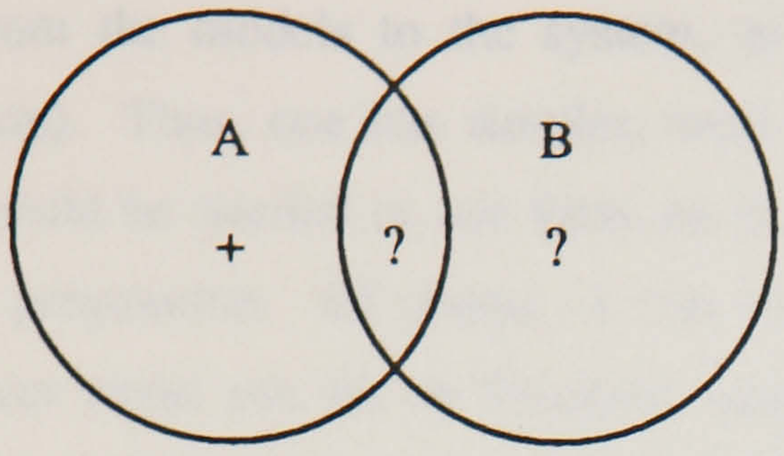
In fact, though (as was drawn to my attention by Barry Richards), it's not obvious how this can be understood in the same way. For instance, in III it's clear that 'All B are A', and Johnson-Laird's model just given doesn't seem to capture that possibility. This

arises from the peculiarities of Johnson-Laird's handling of negation, in which cases it's especially clear that the form of the model depends importantly on the *syntax*, in a certain sense, of the proposition being modelled. In the account of (a), at least, it's evident that 'a ≠ b' indicates the presence of an A equal to no B at all, and this is because (i) it's not equal to that particular B, and (ii) the model-handling procedures are such that it can never be made equal to any B. In the model under discussion, then, the first A cannot subsequently be made to equal the last B. On the other hand, the first B *could* later be equated with some other A. Hence 'a ≠ b' indicates an A which is not any B, but not necessarily vice versa. And thus the model could be 'recursively revised' so that all the Bs were in fact linked to some A, and therefore it is at least *compatible* with III.

Here we have one reason for Johnson-Laird's recognition that 'the content captured in a model is therefore a function of both the model and the processes that can revise and evaluate it' ((b) 35 - but similar remarks occur in several other places throughout Johnson-Laird's work - *cf.* chapter 9, above). We have already looked at this point in some detail, but we should pause here, as it were parenthetically, to note its relevance to the relationship of the theory in (a) to that in (b), as detailed in chapter 8, above. According to Johnson-Laird, these differ 'in several substantial ways' ((b) 6). Certainly, there are some detailed differences in the processes which are supposed to occur, and to some extent in the predictions which these yield (although not the major ones). However, I want to suggest that too much is made of some differences, particularly the accounts of model-building and the figural effect. It seems to me that (b) could be regarded as essentially just adding more detail to the account in (a), which really contains very little. We are told that the arrows show an 'asymmetry' in the *representation* (the mental model) used by the reasoner - but we know little of what this might mean. Is there any clear reason to deny that there is such an asymmetry in (b), bearing in mind the processes in working memory? The point is that Johnson-Laird often draws only a vague distinction between the representation and the mechanisms which manipulate it, and it isn't obvious that the effects of the arrows in (a) shouldn't anyway be ascribed to the latter. The issue is only confused by Johnson-Laird's comparisons (in Johnson-Laird, 1983, esp. ch. 7) of mental models with images, and the like, which are wholly *inert*, in a certain sense, and make it difficult to see how such a relatively active bias could be therein encoded. (Johnson-Laird's ideas about 'list-processing pointers' seem obscure, and too briefly mentioned to be helpful, even where they are introduced in a little more detail in 'Models of Deduction', 1975.)

A major argument of Johnson-Laird's is that his models can each represent the content of a statement, and that this is a significant advantage over the Euler Circle type of theory, where the number of representations needed for a whole syllogism can quickly get out of hand. But obviously one could decide to annotate one's circles in something like the usual Venn diagram fashion, to show whether or not they are filled, and even add a

symbol to indicate areas *optionally* filled. Hence 'Some A are not B' might appear as:



which seems to say all that needs to be said. And, in a sense, it codifies negation quite as explicitly as Johnson-Laird does: this is built into the conventions determining how the model is read.

No model is a model simply in virtue of its structure - what matters is how this determines the way it's *used*. Johnson-Laird's models could be used so as to represent any arbitrary thing, though it might have to be a very odd system which so used them. But what tells us how they're used must be a specification of (i) the features of the representation, and (ii) how the system reacts to these. For a representation of a certain complexity, it seems that the complexity must be distributed between these two, with some sort of 'labelling' system helping to decide how. Johnson-Laird (1983 101), for instance, says that a series of his models could be rolled all into one, so that for the 3AE syllogism we get:

```
c      (= a)
c      (= a)
-------------
   b = a
   b = a
     (a)
```

which carries the penalty that a 'new notational principle' is needed in the interpreting system to treat this as three models were previously treated. But, he says, 'such a device is merely a notational variation...'; what this suggests to me is that, eg, the representation for 'All A are B' is just a notational variation for the *two* models:

```
a = b        a = b
a = b        a = b
                 b
```

(and *mutatis mutandis* for all the others) and consequently there is no reason in principle for the Euler Circle theory to be regarded as inherently more complex. It's just a matter of where the complexity is located. Similarly, one could move the representation of negation from the models to the system, as indeed Robert Inder advocates (personal communication). Thus, one has simpler, more 'natural' models, but a rather complicated arrangement would be needed to use them so as to achieve a complete representation of the meaning of a proposition. Of course, it can well be doubted that any such complete representation is ever made use of, as Erickson implicitly suggests. Johnson-Laird's grounds for rejecting this, based on notions of 'competence' and 'rationality', are suspect indeed, as we found in Part I. And at least as doubtful, perhaps, is the claim that the nature of the trade-off between models and processes can be empirically determined (cf. Anderson, 1978). In this case, the only important difference between the suggested view and Johnson-Laird's is that his employs discrete *tokens* rather than *areas*. We'll look at this point shortly.

11.3 Venn Diagrams

We have just been suggesting that there are few compelling reasons arising from psychological considerations, to suppose that Johnson-Laird's theory is superior to some kind of labelled Euler circle representation. The most obvious such representational system is that provided by Venn diagrams, and there have been theories which proposed cognitive models based on these. Johnson-Laird (1983, ch. 4) devotes particular attention to one offered by Allen Newell (1981), in which the diagrams are actually represented by strings of symbols. The symbols indicate the properties that entities can have, given some premise, as eg. 'All A are B' leads to

Nec A+ B+, Pos A- B+, Pos A- B-

which says that something is necessarily both B and A, and that there are these two other possible combinations. The reader will observe the similarity between this and the 'model classes' described in the last chapter.

Johnson-Laird goes on himself to suggest a simplification of Newell's scheme, in which a 'table of contingencies' can be constructed for premises concerning A, B, and C (not unlike a truth-table) as follows:

A	B	C
+	+	+
+	+	-
+	-	+
+	-	-
-	+	+
-	+	-
-	-	+
-	-	-

'All that is needed to complete the theory', says Johnson-Laird, 'is a procedure that establishes the positive existence of certain contingencies' (92). Assuming that negated features can be treated simply by ignoring them (which allows us to ignore the entire last line of this table), entities falling into these contingencies are evidently just the same as the seven required for the use of Smiley's proof in the last chapter. What we then see is that if the argument of that chapter is correct, Johnson-Laird's theory appears as a system for operating with these contingencies in just the sort of way needed 'to complete the theory'.

Johnson-Laird's criticisms of this approach boil down to two: he points to 'the difficulty of mentally manipulating tables of contingencies', and claims that 'Newell holds fast to logical power, but fails to explain systematic error' (93). On the account we have developed, however, he seems to have solved the first of these problems himself, if somewhat indirectly. And in so doing, he has provided Newell with a means of reformulating a Venn diagram theory which explains all that needs to be explained. The point is that in all formal aspects, leaving aside what can reasonably be attributed to mere 'notational variation', Johnson-Laird's theory itself can be seen as precisely a Venn diagram theory, in which the notation is actually very similar in effect to Newell's, and in which the apparatus of forming the initial sets of contingencies (initial models) through a certain heuristic, provides a simple way of reducing the complexity of the search through the contingency table.

11.4 Natural Models

The salient apparent difference between Johnson-Laird's models and the circle-based ones we have lately been considering, is that his depend on the use of individual tokens to represent entities which stand for a set, while the others notionally use an area (an infinite set of points). This difference, Johnson-Laird claims, gives his models the advantage of being *natural* in a way that the others are not, in so far as their structure 'corresponds directly to the structure of the state of affairs that the discourse describes' (1983 125). This is because his models, like the situations they represent, contain finite sets of related individuals, whereas the others 'map finite sets of individuals into infinities of points'.

There is, even *prima facie*, something curiously implausible about this idea. What is 'unnatural' about picturing sets of entities, whatever their cardinality, as areas in a plane? Areas intuitively share the most important properties of sets (intersection, etc.) and present these in a way which is one simultaneously of striking clarity and obviously 'analogical'. This is much more obvious than is the same fact described in terms of infinite sets of points, and it is also much more true to the way in which Venn diagrams and the like work, since the alternative description of areas as sets of points plays no role at all in their use, as is reflected by the fact that areas are replaced in, eg., Newell's notation, by *single tokens*, not infinities of them. There's no obvious gain over this in Johnson-Laird's move of using arbitrary numbers of tokens (except possibly where numerical quantifiers and similar problematic items are involved, but we are presented with little real evidence that Johnson-Laird's system is in fact any better for these cases). For the domain of syllogisms as such, his argument therefore collapses.

A significant factor here is the notion of representation that is at issue, and Johnson-Laird's ideas about how it works. He asserts that his models are more natural because their structure is more like ('corresponds directly to', he says) the structure of the situation modelled. But there is no clear criterion of 'directness', or resemblance, defined. We saw in chapter 9, above, that modelling depends on the identification of some equivalence relation on situations, and another on models, and a mapping which associates equivalent models with some set of equivalent situations. This is a relatively complex arrangement, but it is nowhere hinted that one can grade these equivalence relations in some way along a scale of 'naturalness', and it is far from clear what the import of it would be if one could. Evidently Johnson-Laird's models can be seen as related to just the same classes of equivalent situations (those in which certain propositions hold true) as are certain Venn diagrams; in the one case, sets of model-tokens relate to sets of entities having some sortal property, in the other areas in a plane relate to the same sets; in both cases, there is a certain well-defined structural relationship between the model and the world, but in neither case is there any way of assessing how 'direct' that relationship is.

There is, perhaps, an intuitive tendency to hold that using, say, Johnson-Laird's models (or Venn diagrams) to represent something radically different from the usual, would be a much less natural way of using them. It might be less natural to use Johnson-Laird's standard model of 'All A are B' for that than for 'Two A are B', and so forth. This only seems compelling when the difference is truly radical, however, and even then it is less so if it is not so much a difference in the use of a certain type of model, as a difference between similar uses of different types of model. Venn diagrams are as natural a way to use circles as Johnson-Laird's models are to use tokens, and the nature of the choice between these is opaque.

If a *natural* model is not one which is differentiated in its formal character, what then is it? The distinction might be argued on psychological grounds, no doubt, but Johnson-Laird makes no attempt so to argue, other than by claiming empirical support for his theory as a whole. Even then, what evidence could establish that subjects were using tokens rather than areas in their models? Presumably, even for the cognitive psychologist, reliance on the introspective phenomenal reports of subjects would be a retrograde step. For what it is worth, however, my colleague Robert Inder (personal communication) once did some syllogistic reasoning experiments, after which subjects were asked what they could tell about how they performed the task. Many of them reported using some form of imagery involving *circles*!

11.5 The Form and Substance of Models

It appears, then, that if Johnson-Laird's theory can be recast as one which uses a quite different type of model (eg. circles), while continuing to work in essentially the same way as a solution procedure, there is no reason to suppose that it is any the less valid in that guise. I have claimed that it can be so recast, and that accordingly the question of what kind of representation reasoners use, if this question is intended to be answered with any psychological exactitude, is left by Johnson-Laird entirely moot. The contact which the theory makes with the evidence comes only through its predictions about the spread of conclusions, the figural effects, and relative difficulties. These predictions can remain unchanged through quite severe redescriptions of the models involved in the theory.

I am inclined to put this point by suggesting that what counts is the *form* of the models, rather than their *content*. What I mean here is that if a procedure involving any kind of model whatever, whether based on tokens, circles or something quite different, can be analysed formally in the above terms of identifying equivalence-classes of situations (ie, if it can be mapped in some way onto the theory described in the previous two chapters), then it is an instantiation of Johnson-Laird's theory considered as a solution-procedure. The latter therefore in an important sense says nothing about the kind of model involved, about those aspects of its structure that do not affect the required mapping.

This criticism is actually even more damaging than it might seem at first sight, since if it is true it shows that Johnson-Laird's theory boils down to the following. The domain of syllogisms can be analysed formally as a set of problems about the relationships of objects falling under three predicates, the complexity of which is expressible in terms of the number of distinct ways there are of representing the premises of a problem in any system having these properties:

(i) all predicates are represented as having non-empty extensions;

(ii) intersections known to be empty are not represented at all;

(iii) other intersections are represented as possibly empty.

The relative difficulty and spread of conclusions can be predicted from the complexity thus measured of each problem, while the figural effects and other less major features of reasoning performance can be derived from assumptions concerning the operation of whatever system constructs and manipulates the representations. (In general, there is little reason to suppose that these assumptions cannot be fairly common to all such systems).

Clearly, Johnson-Laird's own models constitute a representational system of this type, but so do the 'labelled Euler circles' discussed above. Classical Euler circles and Venn diagrams default in various ways from the specified requirements, but this defect can be overcome by adjusting the construction and manipulation procedures, in much the sort of way mentioned by Johnson-Laird when he observed that in principle all problems could be made one-model problems (*loc. cit.*). There are, indeed, many different ways of alternatively characterising essentially the same measure of complexity in syllogisms, any of which would serve as well from the point of view of making the predictions.

11.6 Conclusions

It is certainly a considerable advance on Johnson-Laird's part to have discovered that the psychological difficulty of syllogisms can be predicted on the basis of their formal complexity as just outlined. We can appreciate the way in which this substantially transcends the sort of results obtained by previous investigators, who tended to treat the experimental results as brute facts having no apparent explanation with any real relation to the logic of the problem. And on the other hand, logically-based theories like Newell's are by and large much more remote from experimental data than Johnson-Laird's. There is also much psychological interest in his discussions of the assumptions necessary to account for the figural and other effects. However, we have claimed that Johnson-Laird's theory cannot be supported when it goes further and attempts to found detailed speculation about the nature and structure of 'mental models'.

All this, of course, has been discussed within the context of the domain of syllogisms. Johnson-Laird is entitled to retort that our criticisms are substantially defused when placed against the wider background of his general theory of mental models, which may find much greater support for his detailed suggestions from application of the theory in other domains. In the next chapter, we shall find, among other things, reasons for doubting that this is in fact the case.

12

Models, Logics and Semantics

In this chapter, we shall look at Johnson-Laird's claims for the semantic role of his models, and his criticisms of the role often identified for logic and similar formalisms in other accounts of mental processing. There is much in Johnson-Laird's theory with which I have no quarrel, sharing several of the intuitions upon which it is apparently based. Many of his criticisms of others are sound, but there are many doubtful aspects to what he intends to put in their place. We shall examine these difficulties and afterwards relate them to the preoccupations of the preceding chapters.

12.1 Language Understanding and Representation

Johnson-Laird's main underlying article of faith about theories of linguistic abilities, is that they have to take into account, and explain, the mental representation of the meanings of expressions. Moreover, he insists, 'unless a theory relates language to the world, or to a model of it, it is *not* a complete theory of meaning' (Johnson-Laird 1983, 230). Presumably, the course of using a model as an intermediary step is only acceptable if the theory relates the model in its turn to the world.

Several chapters of the book, *Mental Models*, seem to constitute a kind of distillation of the struggles in an earlier paper (Johnson-Laird 1982), where many topics in the philosophy of language and psychology are embraced in the hope of arriving at a synthesis advanced beyond the previous results of either discipline. Johnson-Laird rejects, in a particularly penetrating discussion (1983 *op. cit.* 191-195), the Kripke-Putnam view that 'meaning just ain't in the head' (which rests on a position of exaggerated Realism), while similarly disdaining the excesses of Psychologism. His conclusion is that language may be related to the world in a substantial variety of ways, reflecting the variety of ways in which it is used, but that these all depend crucially on human cognitive capacity. In fact, it does not really emerge that there is anything identifiable in general as the meaning of an expression. One might put this in a Wittgensteinian way, and say that there are ultimately only

the circumstances of an expression's uses, but that such usage is precisely what has to be accounted for by a psychological theory of linguistic activity. In keeping with the tenor of this characterisation, Johnson-Laird normally puts all his emphasis on internal representations which are not of anything essentially linguistic, but rather of (parts or aspects of) *the world.*

We have already seen this feature of his account in our discussion of his syllogistic theory (chapter 9, above). The models in the array representing what we called the 'expansion' of a given model, are supposed directly to represent states of affairs. Hence Johnson-Laird notes (*op. cit.* 440) that his models are representations, at most, of the *extensions* of expressions. In order to capture the *intension* of an expression, one must look to the processes which construct, revise and evaluate the model, processes which are indeed an indispensable feature of its representative power. In fact, as we have seen, one has to do this even to show how a model represents the complete extension of, eg, a universally quantified expression. This observation somewhat attenuates the role of the model itself in representation - a point we shall return to.

The position here obviously is not unlike that with ordinary first-order model structures in model-theoretic semantics, and in fact Johnson-Laird says that his intention is to elaborate an account whereby he can be seen as explaining *how* the mind can come to compute what model-theoretic semantics says it must compute (Johnson-Laird 1983 167). He remarks that 'mental models are analogous to model structures, but . . . there are important differences between them.' One such difference is held to be in the way basic lexical items are handled. The standard theory says essentially nothing about these, being concerned only with the 'logical content' of an expression. Devices such as 'meaning postulates' are freely employed to respect relations such as synonymy between basic terms, with no attempt to explain these relations. Johnson-Laird notes that in elaborating a theory of mental models, one will have to specify procedures that construct and evaluate models containing entities satisfying particular predicates, and the 'body' of these functions is bound to contain important semantic information about the meanings of basic terms (181). It will be a significant piece of evidence that his models might in fact capture the semantics of expressions, however, if it can be argued that these functions, when specified in a psychological theory, are fulfilling *inter alia* the role picked out in a relatively abstract way by the semanticist's idea of meaning postulates.

The latter are not intended as a psychological theory, and Johnson-Laird is quite justified and highly plausible in his criticisms of psychological theories which have tried simply to adopt them as a mechanism, inasmuch as he argues convincingly that any such approach will either be baroque in its degree of complication, or fail to capture the range of phenomena exhibited in natural language. This in itself does nothing to reduce the

significance of model-theory as a formal account of the semantics of natural language. If, on the other hand, Johnson-Laird is claiming that nothing so inflexible *could* account for these phenomena, the situation becomes more fraught. Having presented himself as accounting for the *how* of computing the *what* which is delineated by formal semantics, he can hardly escape the fact that if meaning postulates will work in specifying the results, they will produce those results computationally if directly implemented (unless thay are not computable), extravagant though it might be to do this. The 'thesis of the autonomy of intensions', as Johnson-Laird characterises these meaning-postulate oriented accounts, does not seem to be in any sense incoherent. Rather, the principal argument is that it is empirically indefensible. Another objection Johnson-Laird has to it is that it is by its nature unable to account for the relationship of language to the world (*op. cit.* 241), but it is unclear that it might not yet do this if its processing model were sufficiently elaborated.

* * *

We are therefore to regard Johnson-Laird's models as in some sense formally equivalent to first-order models. Given that the required sense is suitably wide, this is not difficult in the case of his syllogistic models (although there may emerge substantial difficulties in domains of greater logical richness). We have already done it, in effect, in chapter 10. It can be seen from the above, though, that it always involves a major simplification of his underlying theory. Models are models of entities satisfying particular predicates, and it is clear from the context in which this remark arises that it has substantial significance. In particular, it has *inferential* significance. This, indeed, is how it may be possible on Johnson-Laird's account to explain text and discourse phenomena, and default reasoning. What seems unclear is its likely further role in processes such as syllogistic reasoning. The models we have looked at were impoverished to the point where the satisfaction of a particular predicate was achieved simply by using a specific character to act as the token, but in a fuller exposition of the theory we might expect to see some rich and complex informational structure, which would encode much 'background knowledge' about entities of that kind (although there seems to me then to arise the threat of trouble from what is known in the field of Artificial Intelligence as the 'frame problem', which typically appears when revising a detailed representation of some situation).

In a more general explanation of his position (1983, ch. 15), Johnson-Laird says that he is actually proposing a range of different types of model. He distinguishes in particular two as it were *genera* of mental models, the *physical* and the *conceptual*. The former, relatively close to the products of perception, are supposed to represent various aspects of some perceptible situation in the physical world, whereas the latter abstract from these, principally through the mechanisms of recursive revision, but also through various notational devices intended to handle problems such as negation. The models deployed in the

account of syllogistic reasoning are particularly abstract 'monadic' models (425). Monadic models are ones that represent 'assertions about individuals, their properties, and identities between them', being restricted to handling one-place predicates (identified by the token-character) along with these identities. They appear to be a subclass of 'relational' models, which introduce relations between the tokens in a monadic model, such relations being indicated by lines, arrows, etc.

There are other, still more abstract models, such as 'set-theoretic' models, where the tokens represent sets, and there can also be representations for the abstract properties of, and relations between, sets. It seems clear that these models could be used for solving syllogisms at least as successfully, and perhaps as easily, as monadic models. Presumably, if disposed to argue against this claim, Johnson-Laird might suggest that their use would be less 'natural', in some sense; which looks like a fair point. But equally fair, perhaps, is to suppose that monadic models themselves, as abstract, conceptual models, are still 'unnatural', and that their use might well be a learned skill, not to be expected of subjects unaccustomed to formal reasoning. Johnson-Laird on several occasions alludes to the propensity of reasoners to involve all kinds of real-world default knowledge in their reasoning processes (*cf.* his 'Bambi' story, *op. cit.* 52, and his arguments against 'mental logic', *op. cit.* ch. 3, discussed below), and the only apparent reason for failing to involve it in syllogisms is that he has tried to control out all such content effects in his syllogistic experiments. Is there, though, any ground for the assumption that naive reasoners, when not offered meaningful content, employ a sophisticatedly abstract system otherwise quite foreign to them?

We lately observed that a fuller version of Johnson-Laird's theory might well encompass models containing much more information. Such an elaboration might go some way to accounting for those features of syllogistic reasoning in less well-constrained circumstances than those provided for Johnson-Laird's subjects, in which there seems to be a marked effect of the *content* of the terms in the premises. Evans' (1982 105-111) comprehensive survey of research into this effect reveals an ambivalence in views about its nature, but it is evidently something that requires an explanation. The trouble is, of course, that attempting to explain it reduces the clarity of the logical structure of the syllogistic problem. Various kinds of inferential processes will have to be introduced, other than those specified in Johnson-Laird's algorithm as we have seen it, and these are bound to interact in complex ways with the ones already there. To get round this, it could be argued that the experimental subjects in fact do use richer models like these but that, since there is no effective content, any content they have in their models has no effect; hence their models are equivalent to more abstract ones. In this case, though, there may well be reason to suppose that Johnson-Laird's proffered 'typology' of models is a false one, for what really changes is not the model, but rather the way in which it is used.

Thus to diminish the role of the model, in comparison to the processes that operate with it, is a dangerous course. For it can come to make the model seem peripheral, even epiphenomenal, in the course of events constituting cognition. If a model can represent almost anything, if only used properly, then it *explains* the representation of almost nothing. Explanation will clearly reside rather with the procedures. But Johnson-Laird consistently says very little about these, beyond a brief indication of what they ought to do; he even describes them as 'ineffable' (*op. cit.* 446). Construed like this, therefore, the whole theory of mental models comes desperately close to being what it so earnestly strives not to be, *viz.* a mere specification of what has to be computed, that sheds no light on how it might be done. The question is begged, of how language is related to the world, because it depends wholly on an account of how language is turned into models, and how the models in their turn are related to the world, neither of which steps is helpfully clarified if nothing very substantial can be said about the nature of the models involved, nor about the processes that subserve their use.

What we are left with is the *idea* of a mental model, as opposed to some other kind of internal representation (for instance, one using explicitly sentential logic). If doubt is cast on the value of the detail of the model theory, perhaps there is yet a distinction here that can significantly illuminate the nature of reasoning processes. To examine this question, we should turn our attention to Johnson-Laird's criticism of 'mental logic' theories, and the reasons he finds for erecting the apparatus of mental models in their place.

12.2 Logics in the Mind

We have looked already at the argument about internal-sentence accounts of representation for the general question of understanding language. This is closely related to an argument about reasoning, and whether it requires, or in any event actually uses, some set of explicit logical operations (applications of rules of inference) in conjunction with essentially sentential expressions. Johnson-Laird characterises this latter view as 'postulating that there is a logic in the mind', and as an example of an extreme view quotes Inhelder and Piaget's 'reasoning is nothing more than the propositional calculus itself' (Johnson-Laird 1983, 24). The argument about this view is, it seems to me, somewhat confused, and deserving of effort in its unravelling. Johnson-Laird declares that

> some version of the doctrine appears to have been held by every psychologist who has considered that human beings are capable of rational thought. It is also embroiled in the nineteenth-century claim that the laws of thought are nothing else but the laws of logic. (*loc. cit.*)

He goes on to analyse the defects of the theory, of which he says the most serious is its treatment of reasoning error. Error should not occur, if guided by a logic, and indeed many theorists have supposed that it does not occur *in the reasoning*, but rather in various

ancillary procedures.

This, of course, is one of the issues we discussed in Part I of this thesis. It is always possible to adjust parts of one's theory in order to keep another part unchanged, while accounting for discrepant data; but often only at some cost in plausibility. Johnson-Laird suggests that this cost for mental logic theories is so severe as to have bankrupted them. There are many respects, however, in which his own theory rests on similar assumptions. He criticises Henle (1978), on the grounds that

> She suggests that mistakes arise because people misunderstand or forget premises, and because they import additional and unwarranted factual assumptions into their reasoning. They fail to stick to pure logic, though they are capable of it. . . . I believe that this defence is mistaken. (*op. cit.* 25.)

It is true enough that Johnson-Laird assumes correct understanding of premises, and in his syllogism theory (though not elsewhere) ignores factual influences, but he nonetheless insists that people are capable of sticking to pure logic. It is an article of his faith that reasoning has always the potential for arriving at a valid conclusion - and for the right reasons. Error on his theory is attributed to faulty reasoning no more than it is on Henle's; it is put down to limitations on short-term memory, and its interaction with other unprincipled impedances. He certainly does not propose that people *use* some specific logic, in the sense that they mechanically apply some kind of explicit mental implementation of its rules of inference to explicitly represented sentential premises, but then it isn't clear that Henle does. It is even less clear that most previous psychologists interested in reasoning thought so, nor the 19th century 'laws of thought' camp. Indeed, Johnson-Laird seems to be in spirit firmly among the latter.

There is a good deal of obscurity about what is meant by the notion of 'a logic in the mind'. Johnson-Laird starts off by using this phrase with a rather wide interpretation, but gradually narrows it quite substantially, until it entails a particular kind of cognitive theory. There is obviously a sense in which his theory involves a logic in the mind: he might prefer (say) 'a logic *of* the mind', but then I would argue that this is a better way to present many previous views on the matter. Few indeed of these views have made any real claims about mental *processing*, or even *representation*. Even Piaget, who certainly has been more outspoken than most, is usually vague about this sort of thing. Johnson-Laird quotes the following, as an example of something to which one is undesirably committed by the doctrine of mental logic:

> the subject will ask himself two kinds of questions: (a) whether fact x implies fact y . . . To verify it, he will look in this case to see whether or not there is a counterexample x and non-y. (b) He will also ask whether it is really x which implies y or whether, on the contrary, it is y which implies x . . . (*op. cit.* 34; from Beth and Piaget 1966.)

This, however, says nothing about the processing that realises these internal self-interrogations. It says nothing about the application of rules of inference; it might just as well depend on a grasp of the truth-conditions of the relevant expressions. The description of case (a), in particular, even seems quite a plausible candidate for something done by examining a mental model! Certainly, Piaget apparently leaves room here neither for error nor for the effects of content, but he may be taken to be describing the ideal case. (In subsequent writings, he seems in fact to have taken the matter of content explicitly into account, rather disconcertingly for some of his critics. See Evans, 1982 *op. cit.* 221). Given his own emphasis on competence and rationality, Johnson-Laird will have to agree that in ideal cases subjects must be able to seek these kinds of counterexamples; but then where is the inconsistency with Piaget?

If the reasoning competence of an individual is describable by a logic, then in an important respect it follows that whatever kind of processing realises the relevant behaviour somehow embodies that logic. Johnson-Laird points out that there is a question as to which logic is the logic in question, when discussing mental logic theories, but largely disregards this when formulating his own account of competence. In fact, the notion of valid inference, upon which he bases his own account, cannot be divorced in this way from particular logics; which logic you choose may depend, say, on whether you view the disjunctive syllogism as valid. It is clear enough that Johnson-Laird is a classicist, in that classical logic can be most easily used to characterise the competence he ascribes to reasoners; he also rejects the intuitionistic treatment of the law of excluded middle, for instance (*op. cit.* 188,445). To realise this competence, one might indeed postulate that computational processes exist which in a straightforward way implement some formalisation of classical logic, eg. as axiomatised in a particular way, or as a natural deduction system. (Johnson-Laird speaks as though it makes sense to seek an empirical answer to which of these might be psychologically 'correct'; even if it does, I certainly share his view that none is likely to emerge.) Many of the previous theories he discusses, however, do not address the issue in such a way as clearly to make this kind of assumption; they are implementation-neutral, one might say, and to that extent do not compete with his own account of implementation. These are such as Piaget and even, I claim, Mary Henle.

Henle (1962 *op. cit.*) does speak in suggestive ways, enquiring whether 'the rules of the syllogism describe processes that the mind follows in deductive reasoning' (368); but the terms of her paper invite one to read this as entailing only that the rules parallel the processes in so far as they generate the same conclusions for a given syllogism, or in effect simply that there exist processes *of some kind* realising the competence described by the rules. In the more recent computational vein, though, one does find examples of theories making just the sorts of claims described above (eg. Braine 1978, Rips 1982), and Johnson-Laird's psychological arguments are fairly telling against these, although in ways

that sometimes appear ambivalent to his own position. On the question of how logic is formulated in the mind, he observes that 'it is difficult to obtain direct empirical evidence *relevant to* this issue' (1983 *op. cit.* 39; my italics) - but relevant evidence surely includes any that would tend to show it not to be formulated at all. What this suggests is that when it comes to describing how a particular competence is realised, one can offer a large range of different possible processing models, some of which can be distinguished on the basis of their detailed behavioural predictions, but that it will remain questionable how far these can be further distinguished on logical grounds. The ways in which accounts can be divided up empirically may not be helpfully correlated with the ways they divide up logically; it then becomes unclear how one might find a basis for saying anything at all in detail about the nature of the logic (or lack of it) involved.

Even considered as a 'straw man', the doctrine of mental logic does its job for Johnson-Laird, in showing itself to be psychologically problematic. The question then is whether he has solid ground on which to found his replacement for it. The argument goes that mental logic is unworkable primarily because of its dependence on syntactic rules and representations - because it fails to capture or employ the *semantics* of these things. Much as in the argument about language understanding in the large, Johnson-Laird insists that it is only by building up a description (however abstract) of the situation at issue in the reasoning problem, that the sorts of solutions people come up with, at least, can be accounted for. We'll now look at this part of his theory.

12.3 Semantics and Reasoning

Johnson-Laird's reasons for supposing that mental processing should in some sense make essential use of semantics, are a complex mixture of his dissatisfaction with mental logic theories, as we have just seen it, along with an intuitive conviction that mental models are a much more psychologically compelling approach and are essentially semantic, and a view that there is a problem in accounting for people's tendency to draw only certain of the logically valid conclusions to particular premises. We have already come across this last argument in Part I (chapter 4), where we saw Johnson-Laird attack the idea of competence that apparently comes out of a simplistically logic-based approach. He proposes an algorithm that purportedly uses only the truth-conditions of propositions in drawing a conclusion from a set of premises, and which only draws conclusions that contain more 'semantic information' than the premises, and/or express it more 'parsimoniously' (Johnson-Laird 1983 *op. cit.*, ch. 3 *passim*). The sense in which this actually involves semantics seems merely to be confusing, however, since it consists solely in the procedure's employing a certain symbol, *true*, at a particular stage in a stepwise sequence of rule-applications. As Barry Richards (1986) has pointed out, this symbol could easily be replaced by another, eg. 'p or not p', which looks much less semantic and much more

'logical', without affecting the algorithm. If this were done, we should have an algorithm obviously operating with purely syntactic logical resources, which nonetheless only drew conclusions from the subset defined by Johnson-Laird as appropriate. That this should be possible does not seem surprising, since rules of inference in standard logics have more of a normative intention than a generative one, and there can be many ways of operating in accordance with them, without going through all the combinations they might sanction.

It is not at all clear what Johnson-Laird intends by remarking that the process involves no 'formal rules of inference'. The point seems to be that the algorithm does not invoke explicit representations of logical rules, in the usual sense of modus ponens, disjunctive syllogism, etc.; it does not have the surface appearance of a proof-theoretic operation. In this respect, it is certainly distinguishable from the systems of those such as Braine (1978 *op. cit*) who propose a very explicit representation of a particular logic in their processing. However, this is not enough to make it non-logical, nor in any exceptional sense semantic. In fact, it is not even clear that the surface appearance is of any crucial importance, for assessing these properties. Johnson-Laird *describes* the algorithm in semantically-loaded terms ('the procedure depends on a knowledge of the truth-conditions of the connectives'), but its operation is equally describable in quite different terms. It is, of course, possible to describe a valid inference in both semantic and proof-theoretic terms (provided it belongs to a system which is *complete*, ie. in which all valid consequences are deducible). But also, when Johnson-Laird's algorithm reduces the premises

if p or q then r
p

to

if true or q then r

it is only doing essentially what the resolution-based 'logic-programming' language Prolog (Clocksin and Mellish, 1981) would do with the same problem. And one does not normally encounter claims that resolution theorem-provers, or Prolog, are somehow especially semantic; more likely, resolution will be held to be a rule of inference.

If this is what it does, does it matter *how* it does it? Would it matter, for instance, if Johnson-Laird's program used resolution? This seems to be one of those questions about the *level* at which one chooses to describe something. In analysing the behaviour of an actual computer, for instance, there are always many ways it can be described. If it's running a Prolog program, then one can describe that, or one can describe it in terms of the activity of the Prolog interpreter, or one can talk about the execution of assembly or machine code. These things are all equivalent, in the sense that they describe the same

process, but very different in how they do so, and they are all appropriate for some kinds of interests, but not others. We might analogise Johnson-Laird's project in these terms as attempting to discover the best (most interesting, useful, predictive, etc.) account of mental processing, represented as a 'program' suitably adjusted for application to the brain. In that case, he might be saying that, whereas there may be a syntactic description of what is going on, it is not as good as a semantic one.

There are places where he produces arguments of this kind (see especially 1983 *op. cit.* 149-154), in maintaining that the computability of mental models, which entails the describability in Turing machine terms of any system using them, does not entail that a semantic description of them might not be most appropriate. One could envisage bending much the same argument to the task of showing at a higher level that it may be more profitable to depict what could in principle be completely described as a syntactic-logic system, as a mental-model system. But this is not typically what Johnson-Laird overtly suggests. He more often seems to imply that the behaviour of human reasoners *cannot* be accounted for as using 'mental logic'; the latter is somehow inadequate to the task. In fact, the claim that emerges is that human reasoning is *too rich* for ordinary logic, that the range of inferences *can only* be explained on a semantic basis (*op. cit.* 140-141). But if this is his claim, it has, as we shall now see, some serious consequences.

12.4 Semantics and Computability

Barry Richards (*op. cit.*) has suggested a particularly difficult problem that arises here for Johnson-Laird, related to one we also noticed earlier (in chapter 4). The problem appears just because Johnson-Laird insists that there is a role for semantics in cognition, which could not otherwise be fulfilled by logic, and simultaneously holds the strong position we have seen with respect to rational competence and computability. A theory, he says, is bound *a priori* to be or somehow embody a computable characterisation of some set of valid inferences. But any such theory is bound equally to be, as Richards makes clear, 'susceptible of a complete proof-theoretic formulation'. There can be no independent role for semantics save in those cases where a logic can be shown to be formulable only in semantic terms. This occurs whenever there are 'semantically valid' arguments (ie, those whose premises, when true, lead only to true conclusions) in the system, which lie outside the set of 'syntactically valid' (ie, provable) arguments. Logics with this characteristic are known as *incomplete*: many such cases can be defined, but naturally they are always non-computable.

It looks, therefore, as though Johnson-Laird will have to give something up. He has forced himself into a corner of his own making; the only way out is to dismantle one of his own walls. One course would be to yield on the mental logic issue. He could concede

that any acceptable theory always *might* be formulated as a logic, but that doing so would be too counterintuitive and inevitably baroque. The alternative seems to be to surrender computability. If mental models were not, after all, constrained to be computable, then there would be room for a semantic characterisation of them which indeed could not be encompassed by any proof-theoretic system; and this characterisation could even still be seen as specifying the competence of the reasoner, with all the consequences for rationality that Johnson-Laird wants to draw from this.

Johnson-Laird appears to have so much invested in his commitment to computability, that the wall on that side must have become far too massive a structure for him to tear down; but can mental-model theory survive a weakening of its opposition to proof theory? In fact, I think the proposed dilemma is to some extent a false one: Johnson-Laird should retreat on both fronts, but give up neither. He can do this, the suggestion is, mainly by retracting his excessive pronouncements about competence. If competence is treated more in the sort of way we suggested in Part I, one comes not to expect a given competence theory to extend beyond some particular domain. The way is then opened for one to maintain a view which claims that reasoning behaviour within a particular domain can be explained on a mental-model based account, and that this might be constrained to be computable, even if its mimicability in principle by a logic-based account is thereby entailed. Mental models might yet be able, though, to capture regularities *across domains*, which were forever denied to the purely syntactic (or fully computational) theory. It would, on this view, be conceded that human cognition is not necessarily all explicable in terms of algorithms - a possibility that Johnson-Laird himself several times alludes to. Creating a theory, however, would involve trying to capture small parts of that behaviour within computational models.

What would be, in a sense, the real work in psychology, would then come in trying to fit these fragmentary models together into a general account of cognition. This might well involve several - perhaps many - different accounts of reasoning, even on similar problems, within different domains; it would certainly involve much that could scarcely be called 'reasoning' at all. It would mean pursuing Johnson-Laird's own objective of integrating accounts of reasoning into a more general conception which involves other processes at different levels, such as the processes which resolve anaphoric phenomena and the like in one's reasoning premises, and on the other hand those processes which somehow enable the reflective, conscious, construction of explicit arguments in defence of a conclusion. If Johnson-Laird's account of explanation is right, however, it would also mean abandoning the idea that cognition is ultimately explicable at all.

12.5 Computability and Explanation

In view of this unfortunate prospect, it is worthwhile to offer a few remarks in defence of the idea that an explanatory psychological theory need not be through-and-through computational. Johnson-Laird insists that it should be, that any theory not capable of being cast into the form of an 'effective procedure' for deriving predictions cannot be scientifically explanatory. We do not wish at this point to become embroiled in a technical discussion of the notion of explanation, of the sort with which the philosophy of science literature is replete; we shall be satisfied with a few relatively intuitive remarks such as those offered by Johnson-Laird. His claim (Johnson-Laird 1983, ch. 1) is slightly confusing since, while for the most part he holds that explanation is impossible without an effective characterisation, he asserts at one point that the 'whole theory' need not be thus stated, and that 'formulating a large-scale theory in this way may require too many *ad hoc* decisions to be worth the effort' (8). It seems unclear whether this means that explanation may occur only on a small scale, or that it may sometimes occur in the absence of an effective statement: if the latter, then perhaps Johnson-Laird's view converges on the one I am offering, but this interpretation is denied plausibility by his rapidly following remark that the existence of human abilities not effectively characterisable would present a boundary to knowledge.

What it seems reasonable to me to hold is that a theory can exist, and can make predictions without there being an effective way to derive these predictions, and can still be of scientific value. Whether it counts as *explanatory* is somewhat moot, since explanation is generally to be assessed relatively to the interests of whoever is seeking it. The contention here can be seen as analogous to perceiving value in a logic which is undecidable, and therefore uncomputable. One can see that an argument is valid, and although one can compute no demonstration of this, one can persuade by appeal to the meanings of the expressions involved. Such, indeed, is precisely the method of argument advocated by Johnson-Laird; his error is only to suppose that in general there will be an algorithm for doing it. In scientific endeavours, we may often be in this position of having to rely on a more-or-less intuitive, but nonetheless fairly robust and 'objective' description of something, from which we draw inferences that are by no means deductively impeccable, if that means subject to formal proof procedures. This might amount to saying there is a place for such 'magical' ingredients as human intuition, but it is nonetheless extreme to claim that it dooms theories based on it to being 'vague, confused, and, like mystical doctrines, only properly understood by their proponents' (*loc. cit.*). An apparent merit of his mental model theory was that it might help to show how we are capable of this kind of theorising, but Johnson-Laird seems to be trying to strangle any such potential facility at birth.

Let us ask ourselves what, in fact, Johnson-Laird has gained by applying his criterion in creating his own theory. There are certain clearly algorithmic, because actually computationally implemented, parts of it. We have, for example, the syllogistic inference mechanism, implemented in LISP-80, and a system that does a certain kind of spatial inference, implemented in POP-11. The problem with these, regarded as explanations, is that they are far too simple, in the sense of incomplete. A very great deal of work is required to take either one of these programs and see it as in any way explanatory of a human psychological ability. A central part of the theory is, in each case, rendered utterly perspicuous, but it inheres in a supporting structure of such unplumbed complexity that it is often hard to know what to make of it. We have to take on trust, on its merit as assessed by the 'magical' ingredient of our own intuition, the major part of the account.

This is not intended as a criticism of Johnson-Laird's theory, but only of his proposed schema for evaluating it, by the light of which I am suggesting that it does not shine too brightly. The theory itself is highly interesting, provocative, perceptive, and certainly of scientific value; but in these respects it may be rivalled by others which evidently make no attempt to comply with the criterion of computability. Some of these are such as the great constructions of Piaget and Vygotsky, but consider, as a less well-known example, the theory of reasoning offered by Eugenio Rignano (1923).

Rignano urges upon us the view that reasoning consists solely in the execution of a series of thought-experiments. His arguments are conducted entirely at the intuitive level. He provides an introspective protocol, near enough, of the processes involved in his reasoning that 'in London, which possesses a population larger than the greatest total number of hairs a man can have, people are to be found who have just the same number of hairs' (Rignano 1923 72-3). He says that he imagined the inhabitants of London lined up in order of number of hairs borne until he arrived at one with the greatest number, but then there remained 'in the background', the 'doubles' of those who must have had the same number of hairs. Similarly, he rehearses Taine's proof of the sum of the internal angles of a polygon, indicating that all the steps of the reasoning are nothing but descriptions of imagined operations performed upon some arbitrary polygon. All this is by no means algorithmic, and there is certainly nothing like an effective procedure suggested for replicating the way in which people might perform these mental experiments. Still there remains this central and captivating idea:

> the logical process is nothing else than a series of experiments, all, theoretically at least, capable of being performed, but limited to thought only in order to economise time and energy. The logical process appears then *to be identical with the perceptual reality itself*, operated solely by means of the imagination instead of actually. (Rignano, *op. cit.* 83, original italics.)

There are obvious respects in which this resembles Johnson-Laird's idea. Equally, it

emerges from a wholly different tradition of psychological thought, and is embedded within a theory of 'affectivities' that is quite foreign to modern cognitive psychology. Johnson-Laird's account does not have quite the phenomenological hullabaloo of Rignano's, although the tableau of actors is rather vividly recalled (*cf.* section 9.1, above). However, the point of citing the example was to play upon the similarities for a moment, and ask what, that is really *explanatory*, is gained by Johnson-Laird in casting his related notion in terms of algorithms and effective procedures.

The principle gain is the ability better to relate certain applications of the theory to empirical investigations. Rignano appears to have been a stranger to the laboratory in any case, but it is clear enough that his theory would be very hard to design tests for. On the other hand, this seems no less true for much of Johnson-Laird's theory. Given a suitably circumscribed situation, with suitably chosen subjects, one can predict the conclusions they will give to a suitably circumscribed set of syllogistic premises, but the extension of this to any other situation depends on accepting a fair number of promissory notes. Mental models are supposed to cover propositional reasoning, reasoning with multiple quantifiers, reasoning in cases where there is significant content in the premises, and so forth. Naturally, no effective procedures are specified for making them do these things, but also curiously little seems either to depend on or to entail the necessity that they ever could be. The explicitly algorithmic parts of Johnson-Laird's theory are in fact rather isolated episodes, serving programmatically to illustrate a conception of how the whole might ultimately be, while elsewhere the argument is often as unclear in its implications as anything in Rignano.

The human brain is obviously a highly complex object, which behaves in appropriately complex ways. We do not know whether it works in a wholly computational manner, and if we did, we could not expect to specify all the algorithms it uses. Still, we have found that computational models are of great utility in modelling highly complex processes, and it would surely be absurd not to attempt their application to this subject. We might succeed in limited prediction, as we succeed in limited prediction of the weather. But, much as we continue to explain the behaviour of the atmosphere in terms that do not yield readily to formalisation, do not appear recognisably in computational models (even though they can be in some sense constructed on the output of these), and do their work through suggestive appeal to our prior understanding of analogous terms from mostly unrelated subject areas such as geography (eg. fronts, depressions, troughs, ridges, occlusions), so we may hope to advance in psychological explanation by means other than the invention of algorithms. We do not claim that only those parts of the phenomenon of weather that are amenable to computational modelling, can be explained at all.

The concept of a mental model is a superb example of such a suggestive, analogistic notion, that we can use entirely because of our pre-existing grasp of what it is likely to be, and not because of an explicit formal definition of how it works, which, even were it to be provided, we could barely hope usefully to comprehend. Such is the complexity of the mind that, even if it is in itself computational, we will never have more than comparatively crude models which attempt to express the gross features of its processing. These will, it is surely plausible to suggest, never be explanatory in virtue of their resemblance to the actual processes they are trying to capture. They will be explanatory, if at all, because we can *understand* them, and there is no evidence that, computational devices though we may be, our understanding of computable explanations is any better than our understanding of other kinds.

12.6 Summary

The arguments in this chapter have been involved and perhaps sometimes unclear. This is rather in the nature of the subject matter, where the issues of semantics, logic, computation and explanation are all so closely interwoven that separation, where achieved, seems always artificial and lacking in respect for the complexity of the problems. On this score, my criticisms of Johnson-Laird are probably as open to criticism themselves, but the intention has been to elucidate some of the issues here. There are a few points that should manifest themselves.

The main question was how far Johnson-Laird is justified in making the claims he does about the role of semantics in his theory. The answer was that some of these claims appear to be in conflict with one another, and with other claims he makes. Centrally, there emerges the combination of Johnson-Laird's view that the role of semantics is irreducible to 'mental logic', with his view that reasoning competence has to be equivalent to some (probably classical) normative logic, and his view that all mental processing is computational. There is an apparently irreconcilable clash between that which is irreducibly semantic and that which is wholly computational. The problem with competence would be apparently of no consequence, if one simply supposed competence to be characterised by a proof-theoretically incomplete logic, except that this also clashes with the condition of computability.

Our suggested way out of Johnson-Laird's dilemma was to adopt a less astringent policy with respect to computability. The value of having effectively derivable predictions from a theory is granted, but one cannot expect this to extend beyond a small, localised region of interest such as, paradigmatically, syllogistic reasoning. In general, a theory will at best consist of many such domain-specific explanations linked together into a whole which is far from clearly amenable to algorithmic characterisation. This does not mean

that it cannot be an explanatory theory, nor that it will be forever unsatisfactory from a scientific point of view.

On the way to finding these fundamental difficulties, we noted the extent to which the major explanatory elements in Johnson-Laird's theory are actually located in the 'ineffable' processes that construct and manipulate models. This is really something which has been a recurrent theme throughout our discussion of his theory, and indeed is something to which he himself often refers. A model can represent almost anything, if used correctly; but we never really know what it is so to use it, nor how such use is facilitated. In these circumstances, the resemblance between Johnson-Laird's theory and others making lesser claims to explicitness (such as Rignano's) is accentuated.

This concludes our detailed investigation into Johnson-Laird's recent work, and the remaining task of the thesis is to try to say something a little more general about the problems that have arisen, and about how they relate to the issues raised in Part I.

13

Synoptic Conclusion

It is now time for a brief summing-up of the discussions in this thesis, and an attempt to draw some general conclusions. The thesis falls into two distinct parts, which in fact it might be feasible to treat as quite separate, but which are yet kept together by strong similarities of interest. It is appropriate here to dwell on these and see what has emerged from the consideration, in the second part, of how the more abstract concerns of the first part appear in a particular theory.

13.1 Algorithms, Competence and Rationality

We began by adopting a certain attitude to cognitive psychology, assuming a strongly computational paradigm. This is not merely arbitrary, of course, since many recent psychological theories have adopted a similar attitude, but it is in any event of particular relevance to Johnson-Laird's approach. In an attempt to make sense of the claims often laid, by psychologists, to recognition for making statements of some importance to logical theory, we developed a view on which reasoning norms and empirical theory exist in a relationship of reciprocal influence and determination. We proposed that reasoning theories grow in an environment of *a priori* logical conditions influencing their emerging account of reasoning competence. It is decided by the theorist that people must be capable of certain kinds of inferences, and these abilities become a cornerstone of his account, often maintained even in the face of recalcitrant observations. There is a limit, however, to how far empirical divergence from the predictions of ideal behaviour can be ignored; it eventually destroys the plausibility of the account as a whole. Hence, the psychologist is brought to a different description of the abilities of reasoners. But the state of imbalance between the psychological and the normative which thus appears leads to a pressure upon the latter. The pervasive doctrine that 'ought implies can' comes into effect, and results in a tendency for logical theory to converge towards the predictions of the empirical investigator.

Where a theory is computational, there should be some account of how, in reasoning, particular premises lead to computational processes resulting in particular sorts of conclusions. In the ideal case, the conclusions will be valid. This is just what happens in Johnson-Laird's theory of syllogistic reasoning. One problem with the idea is that it appears to depend on the deduction in question belonging to a logic with a computable proof system, and this in turn requires that the logic is at least complete (ie, all its valid conclusions are provable in it). Some logics (eg. first-order classical predicate logic) have this property, but others (eg. second-order classical logic) do not. In cases of the latter kind, the logic can be described only semantically; why conclusions follow from premises has to be grasped from an appreciation of their meanings, since no purely formal proof is guaranteed to be available. Even where a logic is complete, it may be *undecidable*, in the sense that there is no effective procedure for deciding whether a given statement is one of its theorems. First-order logic is a case in point, where although it is possible to prove any valid argument computationally, there is no guarantee that a procedure searching a counterexample to an invalid one will ever terminate. Here, one is likely to have to resort to semantic methods to evaluate a given argument.

A promise Johnson-Laird seemed to hold out was that mental models could be used to handle cases of this kind, being of their nature semantic representations not unlike the models logicians standardly use, for instance in determining counterexamples. Unfortunately this ambition is discovered to founder on his simultaneous insistence on computability. Starting with the assumption of fully logical competence, Johnson-Laird is committed to his models somehow instantiating a computation of a function from premises to valid conclusions in an appropriate logic. As we have noted, familiar metalogical results from the theory of recursive functions entail that such a thing can exist only for certain constrained types of logical system, which exclude the ones we had hoped for progress on. It is of no avail to argue that the computation in question is somehow at a 'higher level' semantic, and that it should not be read in terms of logical rule-application. These features are of no consequence; if, say, a theory is undecidable, then it can be proved that there is *no* computational procedure for deciding the validity of an argument stated in it.

What might appear a possible escape from this trap is to suppose that there is a procedure which somehow accidentally or coincidentally just happens to compute (perhaps among other things) the required function, although there is no proof to be had that it will do so. One has a description of it in terms of models, their creation and their manipulation, and it simply emerges from this that valid results are computed. This is related to an idea explored by Dennett (1978, ch. 13), in refuting the by now well-known argument originating with J. R. Lucas, which purported to show, on the basis of Goedel's theorem, that people cannot be Turing machines because if they were they could prove their own Goedel sentences, and this is a *reductio ad absurdum*. Dennett's answer is that people can be

interpreted as Turing machines, and as any number of Turing machines simultaneously, in such a way that for a given such interpretation there is always another that circumvents the Goedelian limitations on the first. Nothing is, in and of itself, a realisation of some specific Turing machine rather than another. A similar approach applied to Johnson-Laird's problem would indicate that perhaps, in principle, a computational device could be seen as, say, a Turing machine computing the theorems of some logic, which would have to have a computable proof system, but from some other perspective could be seen as a different Turing machine computing something else which, as it turns out, includes all the theorems of a non-computable logic. A crucial point about this idea, though, is that *when seen from the perspective of the non-computable logic*, the device could *not* be interpreted as computing its theorems, just as Dennett's device, when seen as a particular Turing machine, cannot be regarded as capable of computing *that* machine's Goedel sentence. Accordingly, even if it is possible that a machine should produce the required theorems, it is impossible that one should have a computational characterisation of it which shows that it will do so.

Johnson-Laird is debarred from exploiting the defence just suggested, by his insistence on the notion of a rational procedure. One might empirically discover of a procedure that it is behaving in a logical manner, but his procedure, he says, has to be seen to arrive at its conclusion for given premises *for the right reasons*, and this evidently entails, if among other things, that the procedure can be *proved* to compute just some function describing only valid arguments. Nor can solace be gained from Johnson-Laird's intention to curb the productivity of this kind of computational system by restricting it to produce only a subclass of valid conclusions, *viz.* those identified by his criteria as appropriately informative. He could only appeal usefully to this feature of his system, if it were shown that the criterion isolates a decidable, or at least complete, sublogic; but this is unlikely to begin with, and anyway several of his remarks (in Johnson-Laird 1983, ch. 6) indicate that he would not find it an acceptable restriction. Johnson-Laird's theory is thus crippled by exactly the feature of formal systems which he held to be 'a final and decisive blow to the doctrine of mental logic' (*ibid.* 141). Any logic so powerful that 'it cannot be completely encompassed by formal rules of inference' is also so powerful that it cannot be demonstrably encompassed by any purely algorithmic system.

* * *

In these circumstances, we indicated that Johnson-Laird ought to withdraw certain of his proposals. In particular, he should take account of our view in Part I, that reasoning theories in general, and their associated accounts of competence, are only feasible within carefully defined 'domains'. Within a domain, it may be that one finds the predictive value of a computational theory to be such that it is worth submitting to the constraints it

imposes. One's focus is narrowed to, perhaps, a specific type of task. The possibility seems to be left open, here, of characterising a wider, if vaguer, view of competence across a number of related domains, so as to arrive at something which is not restricted to a particular kind of logical system, but correlatively is not completely describable in computational terms. We advanced, in partial defence of this idea, some remarks directed against Johnson-Laird's charge that no such theory can be properly explanatory of cognitive phenomena. What is needed is roughly what has been proposed (and what Johnson-Laird goes some way in the direction of providing), which is an account that is highly explicit at certain crucial points, but elsewhere depends on a more-or-less intuitive grasp of the cogency of the supposed relations between these.

In thus withdrawing from some of his more exposed ground, Johnson-Laird should not be thought to be sacrificing anything of value. His belief that only processes such as he suggests could account for the rationality of mankind, for our ability, for instance, to invent formal logic, is surely mistaken. That we are rational in the sense basic to intentional description is, we urged in part I, certain in advance of psychology. Rationality in this sense, though, does not entail some in-principle infallibility in making inferences. Rather, it founds the whole system of interactive behaviour, including most importantly in this context linguistic behaviour, that permits us to engage in activites such as logical theorising. These are *reflective* activities, in which we build upon our observations and experiences of our own and others' reasoning behaviour. There is no good reason to suppose that the cognitive processing involved in these activities is more than remotely related to the processing involved in naively, though howsoever selfconsciously, performing tasks like solving syllogisms (*cf.* chapter 2). There might be insights to be gained here about language-understanding, say, but considered specifically as reasoning tasks we might be arbitrarily bad at such things (and there's no shortage of evidence that we are), while at another level we can come to realise that this is the case, and understand in what ways. This is not to imply even that we can improve our normal behaviour, since in general there is not the opportunity to carry out a complex reflective analysis of a problem before solving it, and the ways of doing so are anyway often vague or indeterminate.

Logical theorising, on our account, is more strongly related to other kinds of theorising than has often been supposed. Johnson-Laird, we noticed, criticises the 19th. century philosophers who regarded logic as a statement of the 'laws of thought', but then himself establishes logic in essentially that position. Our view is opposed to this, if by 'laws of thought' is meant eternally *a priori* rules according to which the mind operates. We accept a view of logic as in some sense aspiring to capture the gradually revealed natural laws of reasoning, but only insofar as we accept the idea that psychology, like other sciences, gradually reveals natural laws. We may then hope eventually to find natural laws about how we are able to find natural laws, but these will be far from imposing Johnson-Laird's

condition on reasoning theories, that they must allow the attaining, in ideal circumstances, of a valid conclusion in all cases, on all tasks.

* * *

Naturally, we want to point to the significance of the claims made here, for the cognitive psychology of reasoning in general, and the major issues of Part I. In domains as rich as normal human situations are likely to be, one essentially cannot expect to have a useful computational account of reasoning which captures a logically-defined notion of rationality. If rationality is adherence to some set of logical laws, then people are at best contingently or accidentally rational, in that it has to be empirically discovered that their behaviour is thus characterisable, and there cannot then be an algorithm predicting it. A competence theory for a domain as rich as this cannot be expressed as the algorithmically-produced ideal behaviour of a reasoner; if the required competence can be part of such ideal behaviour, it cannot be described in the terms of the algorithmic account.

This leads leads us to generalise the suggestion that competence should not be seen as a specification for an algorithm (*cf.* also chapter 4), but more as some sort of ideal to be aimed at by an account of behaviour which will not be computational overall, but which may involve computationally explicit sections directed at particular sets of data. These sections are not then bound by any necessity to predict fully competent behaviour as an ideal for the relevant set of data. In cases where the logical account would be decidable (which holds for syllogisms, as a subset of monadic predicate calculus), one could without difficulty take such a goal on board; but one should not have to. If no subjects reach a particular conclusion on some problem, then there is no reason to build it *a priori* into one's account that they can. There may be good reasons to build it in nonetheless - eg. because it makes the theory as a whole simpler, more extendible, less *ad hoc* - but it is important to distinguish these reasons from the reason that one is striving to capture some notion of rational competence. That notion becomes more of a background, regulative notion, accrediting the aspirations of one's theory at a higher level, specifying the kinds of abilities one wants ultimately to predict. This seems in tune with its role as we saw it in its interaction with normative logical theory.

13.2 Models and Representation

It is also appropriate now to survey our more specific conclusions about Johnson-Laird's work, particularly his theory of mental models, the role it might play in cognitive explanation, and the problems that arise with it, other than those just summarised.

In, initially, showing that Johnson-Laird's theory is in fact a theory of the kind discussed in Part I, we engaged in a fairly extensive analysis of his algorithm for syllogistic reasoning. This was shown to be an effective procedure for deriving valid conclusions to

syllogistic problems (or discovering that there is none), by treating the models as sets of entities with properties, interpretable under some appropriate procedure as representative of particular equivalence classes of such sets of entities, in virtue of their capacity for being 'expanded' into any equivalent set. The expansion procedures, however, while being crucial to the account of the models as semantic constructions, play no role in the algorithm, which instead invokes procedures for deriving propositions from models, as conclusions to syllogisms, in ways that can be interpreted as identifying the important relationships (eg. intersection, inclusion) shown to exist between the extensions of the predicates involved.

In describing these structures and operations, we have, like Johnson-Laird, used thoroughly semantic terminology. We could as easily have used some neutral description of the sets as sets of symbols, and simply shown that the inputs and outputs can be interpreted so as to correspond to syllogistic problems. The question therefore arises as to what lies behind Johnson-Laird's claim that these representations are somehow essentially semantic, rather than logical and syntactic. This brings us back briefly to the points made above about computable systems: if anything is computable, then there is a description of it that is couched in terms only of the activities of a simple mechanism, under which it is no more semantic (and no less) than any other computation. There's a good deal of discussion in the cognitive science literature about how far and in what ways these sorts of descriptions are semantic (good examples are Fodor, 1980, and Smith, 1982), but this is largely peripheral to the present question, which is what justifies a much higher-level description of a system. Can a particular description be somehow the 'right' or 'true' one?

Our answer to this question is essentially that evaluation of such a description has to be on pragmatic grounds, depending on whether it best serves the interests of an explanatory theory. Johnson-Laird might well be justified, for instance, in pointing to the superiority on this score of his theory over 'mental logic'. But there is no answer to the question of which is in some objective sense *correct.* This is not an application of Anderson's (1978) 'mimicry theorem'; rather, it depends on the view that any description of a natural computational system at a level above the purely physical, is entirely indeterminate and depends on some essentially arbitrary interpretation of the operations of the system. The resemblance between this remark and Dennett's argument referred to above is not quite accidental; we concede that the view advocated is seriously infected with a Dennettian instrumentalism. We make no apology for this, however.

The way to proceed is, on almost any account, to use experimentally collected data on the behaviour of a system, to decide the plausibility of theories based on different proposed structures and processes within. We therefore investigated whether Johnson-Laird's proposed models gain anything in predictivity, etc., over certain others from which he has, often vigorously, dissociated himself, although they retain many of the properties seen as

desirable in his own models, such as an evidently semantic role and an 'analogical' structure. We found that, in fact, the features of his models that are effective in his use of them, especially in the syllogistic algorithm, are ones shared by a large range of quite different kinds of models (and other representations), and that there is little reason to suppose that the remaining details are of great relevance to the experimental data. Hence there is no clear empirical foundation for his claims about these details, and nor did we find persuasive his models' characterisation as somehow more 'natural' than the others.

We also separately considered the internal role, in the theory, of the 'typology' of models that Johnson-Laird offers, in that we were interested in the question of how much information the models contain. What emerged here was the overwhelming importance of the procedures that construct and evaluate models. It appears that it is really these which distinguish between different types of model; it is more that there are different types of *uses of* models. Models themselves continue to have a highly abstract characterisation and role, which amounts mainly to a description of certain formal features of the information represented, which are supposed to be paramount in processing it. This is revealing, especially from a logical point of view, but it places a considerable stress on the explanatory importance of an account of the procedures, and none is forthcoming. To describe them as 'ineffable' seems to be to give up hope of a clear account of what is happening in cognition; if it amounts to a recommendation to treat them as computationally primitive (part of what Pylyshyn would call the 'functional architecture' of the mind), in a wholly algorithmic explanation, then this seems rather worryingly to exhibit the potential poverty of the latter. Explanations in terms of algorithms running on extremely complex, but cognitively opaque, machines, seem to leave much to be desired. On the other hand, our suggestion has been that the idea is sensible in principle, so long as it is conceded that one can have cognitively explanatory, but *non*-algorithmic accounts of important features of the machine architecture.

* * *

It would be a mistake to take these arguments as adding up to a dismissal of Johnson-Laird's theory. The idea of mental models, that cognition depends on representations of aspects of the world inhabited by the mind, is one of great power and potential in cognitive science. We have dwelt rather upon the negative aspects of it, partly because our particular interest in reasoning led us to them, and partly because in a somewhat Popperian fashion we regard the best service one can do a theory as being the attempt to destroy it. We have not offered an explicit alternative because it seemed more important to go into the issues underlying the creation of any such theory. If, as argued, the bases of certain inadequacies in this, one of the most recent of theories of reasoning, turn on fundamental considerations for the relation of logic to psychology, the best hope for a useful contribution is

to shed some light on these. This, we hope it will be felt we have done. If we have succeeded in exposing serious difficulties in Johnson-Laird's formulation and defence of his ideas, the consequence should be repair and consolidation, rather than abandonment.

13.3 Reasoning and Logic

Our final message can essentially be summed up as the view that several of the ambitions people have had in relating psychology and logic, have been either misplaced or premature. Reasoning tasks are certainly characterisable in some sense as logical tasks, in which premises and conclusions can be discerned; the extent and nature of the relevance of this to the processing that realises reasoning behaviour is, however, a question of profound complexity. Psychologists have been too quick to latch onto ideas in logic, ideas about competence and rationality, and try to cast their theories in ways that offer insights into these matters. Sometimes, the relationship between the nature of the proposed processes (eg. as algorithmic), and the domain of logic, has been misunderstood. Sometimes, the presuppositions of the theories (eg. that competence must be describable precisely by some well-known logic) have been excessive. Sometimes, the interpretations of the data (eg. as showing that people are thoroughly and irrevocably irrational) have been extreme. Our conclusion is that, for one thing, several foundational questions about rationality and its relations both with logic and psychology need further clarification, and that, for another, there is a delicate balance to be achieved by any cognitive theory that aspires both to be explanatory about mental processing and at the same time carry serious consequences for logical issues. The latter of these goals demands a simplicity and clarity that seems hard to reconcile with the complex subtlety of the mind, while the former is in the opposite position. We do not yet know how, or if, these objectives can jointly be satisfied.

Appendix

Some Syllogistic Reasoning Data Re-Analysed

The following two tables relate to the argument in section 8.4 of the main text. They present a re-analysis of some of the data which appear in Johnson-Laird and Steedman (1978), for the second experiment, second test.

Table 1

Each cell shows, for a given figure, the percentage of conclusions given which could have been read off the IR of a Johnson-Laird model. ('A' shows cases where there is an A premise, 'no A' where there is not.) Notice that this is always small if there is no A premise; and of course it is always also incorrect in such circumstances. Most incorrect IR conclusions occur in the first two figures.

Figure of Premises

	1		2		3		4	
	A	no A	A	no A	A	no A	A	no A
Overall	76	13	74	14	55	14	67	9
Correct	40	0	36	0	35	0	52	0
Incorrect	36	13	38	14	20	14	15	9

Table 2

Form of the First Premise

Form of the Second Premise	A				I				E				O			
Figure of Premises	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
A	A(I)+ 18 *	A(I)+ 19 *	NVC- 8	I- 13	I+ 19 *	NVC- 6 *	NVC- 9 ?	I+ 20 *	O- 3 *	E(O)+ 16 *	E(O)+ 18 *	O- 7	NVC- 9 ?	NVC- 3 *	O+ 5 *	O+ 14 *
I	NVC- 8 *	I+ 18 *	NVC- 15	I+ 20 *	NVC- 16	NVC- 15	NVC- 17	NVC- 18	O- 8	O- 14	O- 13	O- 13	NVC- 16	NVC- 16	NVC- 14	NVC- 19
E	E(O)+ 18 *	O- 0 *	E(O)+ 15 *	O- 7	O- 17	O- 5	O- 7	O- 14	NVC- 13	NVC- 14	NVC- 17	NVC- 13	NVC- 14	NVC- 17	NVC- 14	NVC- 14
O	NVC- 5 *	NVC- 7 *	O+ 14 *	O+ 17 *	NVC- 16	NVC- 18	NVC- 18	NVC- 19	NVC- 16	NVC- 15	NVC- 17	NVC- 14	NVC- 19	NVC- 16	NVC- 20	NVC- 20

Key to each cell:

Top: Mood of correct conclusion. Bracketed are alternative, equally valid conclusions. '+' means the correct conclusion is an IR conclusion; '-' means it isn't. 'NVC' means there is no valid conclusion.

Centre: Number of subjects giving *correct* conclusion.

Bottom: '*' = more than 50% of subjects give an IR conclusion;
'?' = this is a marginal result.

References

Anderson, A R and N D Belnap, *Entailment (Vol. 1)*, Princeton University Press, Princeton, 1975.

Anderson, J R, "Arguments Concerning Representations for Mental Imagery," *Psychological Review*, vol. 85, pp. 249-77, 1978.

Beth, E W and J Piaget, *Mathematical Epistemology and Psychology*, Reidel, Dordrecht, 1966.

Braine, M D S, "On the Relation Between the Natural Logic of Reasoning and Standard Logic," *Psychological Review*, vol. 85, pp. 395-416, 1978.

Braine, M D S and D P O'Brien, "Categorical Syllogisms: a reconciliation of mental models and inference schemas," , New York University, (Unpublished paper), 1983.

Chapman, I J Chapman and J P, "Atmosphere Effect Re-examined," *Journal of Experimental Psychology*, vol. 58, pp. 220-6, 1959.

Church, A, *Introduction to Mathematical Logic, Volume 1*, Princeton University Press, Princeton, NJ, 1956.

Churchland, P, *Scientific Realism and the Plasticity of Mind*, Cambridge University Press, Cambridge, 1979.

Clocksin, W F and C S Mellish, *Programming in Prolog*, Springer Verlag, Berlin, 1981.

Cohen, L J, *The Implications of Induction*, Methuen, London, 1970.

Cohen, L J, "On the psychology of prediction: Whose is the fallacy?," *Cognition*, vol. 7, pp. 385-407, 1979.

Cohen, L J, "Whose is the fallacy? A rejoinder to Daniel Kahneman and Amos Tversky," *Cognition*, vol. 8, pp. 89-92, 1980.

Cohen, L J, "Can human irrationality be experimentally demonstrated?," *The Behavioural and Brain Sciences*, vol. 4, pp. 317-370, Cambridge University Press, 1981.

Cohen, M and E Nagel, *Introduction to Logic and Scientific Method*, RKP (Complete edition, reprinted), 1972.

Conee, E and R Feldman, "Discussion: Stich and Nisbett on Justifying Inference Rules," *Philosophy of Science*, vol. 50, pp. 326-31, 1983.

Coppee, H, *Elements of Logic*, J H Butler & Company, Philadelphia, 1874. (Revised Edition.)

Daniels, N, "Wide Reflective Equilibrium and Acceptance Theory in Ethics," *Journal of Philosophy*, vol. 76, pp. 256-282, 1979.

Davidson, D, "Actions, Reasons and Causes," *Journal of Philosophy*, vol. 60, pp. 685-700, 1963.

Davidson, D, "On the Very Idea of a Conceptual Scheme," *Proceedings and Addresses of the American Philosophical Association*, vol. 47, pp. 5-20, 1973/4.

Davidson, D, "Psychology as Philosophy," in *Essays on Actions and Events*, Clarendon Press, Oxford, 1980.

Davidson, D, "Rational Animals," *Dialectica*, vol. 36, no. 4, 1982.

Dennett, D C, *Brainstorms: Philosophical Essays on Mind and Psychology*, Harvester Press, Brighton, Sussex, 1978.

Dennett, D C, "Three Kinds of Intentional Psychology," in *Reduction, Time and Reality*, ed. Richard Healey, pp. 37-61, Cambridge University Press, Cambridge, 1981a.

Dennett, D C, "True Believers: the intentional strategy and why it works," in *Scientific Explanation*, ed. A F Heath, pp. 53-75, Clarendon Press, Oxford, 1981b.

Dennett, D C, "Making Sense of Ourselves," in *Mind, Brain and Function*, ed. J I Biro and R W Shahan, pp. 63-81, Harvester Press, Brighton, Sussex, 1982a.

Dennett, D C, "How to Study Human Consciousness Empirically," *Synthese*, vol. 53, pp. 159-180, 1982b.

Dennett, D C, "Beyond Belief," in *Thought and Object*, ed. Andrew Woodfield, Oxford University Press, Oxford, 1982c.

Dreyfus, H, *What Computers Can't Do: A Critique of Artificial Reason*, Harper & Row, New York, 1972.

Egan, D E, "Retrospective Reports Reveal Differences in People's Reasoning," *The Bell System Technical Journal*, vol. 62, no. 6, pp. 1675-1697, 1983.

Einhorn, H and R M Hogarth, "Behavioural Decision Theory: Processes of Judgement and Choice," *Annual Review of Psychology*, vol. 32, pp. 53-88, 1981.

Ellis, B, *Rational Belief Systems (APQ Library of Philosophy)*, Basil Blackwell, Oxford, 1979.

Elster, J, "Rationality," in *Contemporary Philosophy, A New Survey (Vol. 2)*, ed. Guttorm Floistad, pp. 111-131, Nijhoff, The Hague, 1982.

Erickson, J R, "A set analysis theory of behaviour in formal syllogistic reasoning tasks," in *Loyola Symposium on Cognition (Vol. 2)*, ed. R Solso, Lawrence Erlbaum, Hillsdale, NJ, 1974.

Evans, J St B T, *The Psychology of Deductive Reasoning*, Routledge and Kegan Paul, London, 1982.

Fodor, J A, *The Language of Thought*, Crowell, New York, 1975.

Fodor, J A, "Methodological Solipsism Considered as a Research Strategy in Cognitive Psychology," *The Behavioural and Brain Sciences*, vol. 3, no. 1, pp. 63-109, 1980.

Follesdal, D, "The Status of Rationality Assumptions in Interpretation and the Explanation of Action," *Dialectica*, vol. 36, no. 4, pp. 301-316, 1982.

Geraets, T F (ed), *Rationality Today*, The University of Ottawa Press, Ottawa, 1979.

Goldman, A I, "Epistemics: the regulative theory of cognition," *Journal of Philosophy*, vol. 75, pp. 509-523, 1978.

Goodman, N, *Fact, Fiction and Forecast*, Bobbs-Merrill, Indianapolis, 1965. (2nd. Edition.)

Grice, P, "Logic and Conversation," in *Studies in Syntax. Vol. 3: Speech Acts*, ed. P Cole and J L Morgan, Academic Press, New York, 1975.

Guyote, M J and R J Sternberg, "A transitive-chain theory of syllogistic reasoning," *Cognitive Psychology*, vol. 13, pp. 461-525, 1981.

Haugeland, J, "The Nature and Plausibility of Cognitivism," *The Behavioural and Brain Sciences*, vol. 2, pp. 215-260, 1978.

Henle, M, "On the Relation Between Logic and Thinking," *Psychological Review*, vol. 69, pp. 366-78, 1962.

Hunter, I M L, "The solving of three term series problems," *British Journal of Psychology*, vol. 48, pp. 286-98, 1957.

Johnson-Laird, P N, "Models of Deduction," in *Reasoning: Representation and Process in Children and Adults*, ed. R J Falmagne, pp. 7-54, Lawrence Erlbaum, Hillsdale, NJ, 1975.

Johnson-Laird, P N and M J Steedman, "The Psychology of Syllogisms," *Cognitive Psychology*, vol. 10, pp. 64-99, 1978.

Johnson-Laird, P N, "Formal Semantics and the Psychology of Meaning," in *Processes, Beliefs and Questions*, ed. S Peters and E Saarinen, Reidel, Dordrecht, 1982.

Johnson-Laird, P N, *Mental Models: towards a cognitive science of language, inference, and consciousness*, Cambridge University Press, Cambridge, 1983.

Johnson-Laird, P N and B Bara, *Syllogistic Reasoning*, September 1983. (Pre-publication Mimeo)

Kahneman, D and A Tversky, "On the Psychology of Prediction," *Psychological Review*, vol. 80, pp. 237-51, 1973.

Kahneman, D and A Tversky, "Subjective Probability: A Judgement of Representativeness," in *The Concept of Probability in Psychological Experiments*, ed. C A S Stael von Holstein, Reidel, Dordrecht, 1974.

Kahneman, D and A Tversky, "On the Interpretation of Intuitive Probability: A Reply to Jonathan Cohen," *Cognition*, vol. 7, pp. 409-11, 1979.

Kekes, J, *A Justification of Rationality*, State University of New York Press, Albany, NY, 1976.

Kyburg, H E, "Recent Work in Inductive Logic," in *Recent Work in Philosophy (APQ Library of Philosophy)*, ed. K G Lucey and T R Machan, pp. 87-150, Rowman and Allanheld, Totowa, NJ, 1983.

Lindemann, E, *A Rhetoric for Writing Teachers*, Oxford University Press, New York, 1982.

Lycan, W G, "Form, Function and Feel," *Journal of Philosophy*, vol. 78, 1981.

March, J G, "Bounded Rationality, Ambiguity, and the Engineering of Choice," *Bell Journal of Economics*, vol. 9, pp. 587-608, 1978.

Newell, A, "Reasoning, problem solving and decision processes: the problem space as a fundamental category," in *Attention and Performance, Vol. 8*, ed. R Nickerson, Lawrence Erlbaum, Hillsdale, NJ, 1981.

Nisbett, R and L D Ross, *Human Inference: Strategies and Shortcomings*, Prentice-Hall, Englewood Cliffs, NJ, 1980.

Nisbett, R E and E Borgida, "Attribution and the psychology of prediction," *Journal of Personal and Social Psychology*, vol. 32, pp. 932-43, 1975.

Nisbett, R E and T Wilson, "Telling More Than We Can Know: Verbal Reports on Mental Processes," *Psychological Review*, vol. 84, pp. 231-259, 1977.

Pylyshyn, Z W, "The Role of Competence Theories in Cognitive Psychology," *Journal of Psycholinguistic Research*, vol. 2, no. 1, pp. 21-50, 1973.

Pylyshyn, Z W, "Computational Models and Empirical Constraints," *The Behavioural and Brain Sciences*, vol. 1, pp. 93-127, 1978.

Pylyshyn, Z W, "Computation and Cognition: Issues in the Foundations of Cognitive Science," *The Behavioural and Brain Sciences*, vol. 3, pp. 111-169, 1980.

Quine, W V, *Word and Object*, MIT Press, Cambridge, Mass., 1960.

Rawls, J, *A Theory of Justice*, Harvard University Press, Cambridge, Mass., 1971.

Rescher, N, *Methodological Pragmatism*, New York University Press, New York, 1977.

Richards, E B, "Review of Johnson-Laird: *Mental Models*," , University of Edinburgh, 1986. (Revised.)

Rignano, E, *The Psychology of Reasoning*, Kegan Paul, Trench, Trubner & Co., London, 1923.

Rips, L J, "Cognitive Processes in Propositional Reasoning," , University of Chicago, Unpublished paper, 1982.

Ryle, G, *The Concept of Mind*, Hutchinson, 1949.

Smedslund, J, "The concept of correlation in adults," *Scandinavian Journal of Psychology*, vol. 4, pp. 165-73, 1963.

Smiley, T J, "Syllogism and Quantification," *Journal of Symbolic Logic*, vol. 27, pp. 58-72, 1962.

Smiley, T J, "What is a Syllogism?," *Journal of Philosophical Logic*, vol. 2, pp. 136-154, 1973.

Smith, B C, *Reflection and Semantics in a Procedural Language*, MIT Laboratory for Computer Science, Report MIT-TR-272, 1982.

Stich, S and R E Nisbett, "Justification and the Psychology of Human Reasoning," *Philosophy of Science*, vol. 47, pp. 188-202, 1980.

Stich, S P, *From Folk Psychology to Cognitive Science*, MIT Press, Cambridge, Mass., 1983.

Stich, S P, "Could Man be an Irrational Animal?," *Synthese*, vol. 64, pp. 115-135, Reidel, 1985.

Stich, Stephen P, "Dennett on Intentional Systems," in *Mind, Brain and Function*, ed. J I Biro and R W Shahan, pp. 39-63, Harvester Press, Brighton, Sussex, 1982.

Thagard, P, "From the Descriptive to the Normative in Psychology and Logic," *Philosophy of Science*, vol. 49, pp. 24-42, 1982.

Thagard, P and R E Nisbett, "Rationality and Charity," *Philosophy of Science*, vol. 50, pp. 250-267, 1983.

Tversky, A and D Kahneman, "Availability: A heuristic for judging frequency and probability," *Cognitive Psychology*, vol. 5, pp. 207-32, 1973.

Wason, P and P N Johnson-Laird, *Psychology of Reasoning: Structure and Content*, Harvard University Press, Cambridge, Mass., 1972.

Wason, P C, "Reasoning," in *New Horizons in Psychology I*, ed. B M Foss, Penguin, Harmondsworth, 1966.

Woodworth, R S and S B Sells, "An Atmosphere Effect in Formal Syllogistic Reasoning," *Journal of Experimental Psychology*, vol. 18, pp. 451-60, 1935.