Pollard, P. and Evans, J. St B.T. (1980) 'The influence of logic on conditional reasoning performance', *Quarterly Journal of Experimental Psychology* 32: 605–24.
Popper, K.R. (1959) *The Logic of Scientific Discovery*, London: Hutchinson.
—— (1962) *Conjectures and Refutations*, London: Hutchinson.
Rips, L.J. (1983) 'Cognitive processes in propositional reasoning', *Psychological Review* 90: 38–71.
Rumain, B., Connell, J., and Braine, M.D.S. (1983) 'Conversational comprehension processes are responsible for reasoning fallacies in children as well as adults', *Developmental Pscyhology* 19: 471–81.
Rumelhart, D., Smolensky, P., McClelland, J.L., and Hinton, G.E. (1986) 'Schemata and sequential thought processes in PDP models', in J.M. McClelland and D. Rumelhart (eds) *Parallel Distributed Processing: Explorations in Microstructure of Cognition*, Cambridge, MA: MIT Press.
Slovic, P., Fischhoff, B., and Lichtenstein, S. (1977) 'Behavioral decision theory', *Annual Review of Psychology* 228: 1–39.
Smedslund, J. (1970) 'On the circular relation between logic and understanding', *Scandinavian Journal of Psychology* 11: 217–19.
—— (1990) 'A critique of Tversky and Kahneman's distinction between fallacy and misunderstanding', *Scandinavian Journal of Psychology* 31: 110–20.
Sperber, D. and Wilson, D. (1986) *Relevance*, Oxford: Blackwell.
Tversky, A. and Kahneman, D. (1974) 'Judgement under uncertainty: heuristics and biases', *Science* 185: 1124–31.
—— (1983) 'Extensional vs intuitive reasoning: the conjunction fallacy in probability judgment', *Psychological Review* 90: 293–315.
von Neumann, J. and Morgenstern, O. (1947) *Theory of Games and Economic Behavior*, Princeton, NJ: Princeton University Press.
von Winterfeldt, D. and Edwards, W. (1986) *Decision Analysis and Behavioural Research*, Cambridge, Cambridge University Press.
Wason, P.C. (1960) 'On the failure to eliminate hypotheses in a conceptual task', *Quarterly Journal of Experimental Psychology* 12: 129–40.
—— (1966) 'Reasoning', in B.M. Foss (ed.) *New Horizons in Psychology*, vol. 1, Harmondsworth: Penguin.
Wason, P.C. and Johnson-Laird, P.N. (1972) *Psychology of Reasoning: Structure and Content*, London: Batsford.
Wildman, T.M. and Fletcher, H.J. (1977) 'Developmental increases and decreases in solutions of conditional syllogism problems', *Developmental Psychology* 13: 630–6.

Chapter 2

Reasoning theories and bounded rationality

M. Oaksford and N. Chater

INTRODUCTION

In this chapter we will argue that considerations of bounded rationality may fundamentally alter our present conception of the adequacy of psychological theories of reasoning. Since its inception cognitive science has been concerned with the limitations on the cognitive system which inhere in virtue of the organization of human memory and the need to act rapidly in real time (Simon, 1969; Kahneman *et al.*, 1982). Simon (quoted in Baars, 1986: 363–4), for example, says that: 'cognitive limitations have been a central theme in almost all of the theorizing I've done. . . . They are . . . very important limitations on human rationality, particularly if the rationality has to be exercised in a face-to-face real-time context'. Cognitive limitations mean that people may be incapable of living up to normative but computationally expensive accounts of their inferential behaviour,[1] i.e. human rationality is *bounded*.

The two most important limitative findings of cognitive science both affect human memory. The constraints imposed by people's limited short-term memory capacity have been mapped out in some detail (Miller, 1956; Baddeley, 1986) and have been appealed to in order to explain certain biases in reasoning experiments (Evans, 1983a; Johnson-Laird, 1983). Perhaps a less-well-known limitative finding applies to retrieval from long-term memory.

In artificial intelligence this limitation has been labelled the *frame problem* (McCarthy and Hayes, 1969; see Pylyshyn, 1987 for overviews). This term tends to be used generically to describe a cluster of related problems, which as Glymour observes, are all of the following form: 'Given an enormous amount of stuff, and some task to be done using some of the stuff, what is the *relevant stuff* for the task?' (Glymour, 1987: 65). Some variant of the frame problem may arise for any task requiring the deployment of prior world knowledge. In this chapter we will trace out the consequences of the frame problem for theories of reasoning. We will argue that a bounded-rationality assumption may have to be made in

deductive-reasoning research, just as in research into risky decision making (Kahneman *et al.*, 1982).

We begin by outlining the range of contemporary theoretical approaches to reasoning based on the taxonomy provided by Evans (1991) and suggest that bounded rationality provides an additional criterion of theory preference. We then introduce an important and implicit assumption which motivates interest in these theories. This we have called the *generalization assumption* (Oaksford and Chater, 1992). It states that theories of reasoning developed to account for explicit inference in laboratory reasoning tasks should generalize to provide accounts of other inferential processes. We will also offer a general characterization of these inferential processes. We then outline more precisely how the limitations of the cognitive system may militate against certain process accounts by briefly introducing *computational complexity theory*. We will then show how complexity issues have raised problems for theories of perception and risky decision making and for theories of knowledge representation in artificial intelligence (AI). We then argue that contemporary reasoning theories are all likely to fall foul of the same problems. We therefore conclude that these theories are unlikely to be psychologically real.

An important corollary to this argument is that because our reasoning abilities are bounded, empirically observed deviations from optimal rationality need raise few questions over our rationality in practice. The interesting questions are how rational the system needs to be to qualify as a cognitive system (Cherniak, 1986), and what kind of mechanism needs to be postulated to implement it (see, for example, Levesque, 1988). To end on a positive note, therefore, we will suggest that, following Rumelhart *et al.* (1986) and Rumelhart (1989), recent advances in neural computation may suggest mechanisms which more adequately address the issues we raise in this chapter. We will also suggest some ways in which reasoning research may develop profitably in the future to identify the kind of rational mechanism (Fodor, 1987) people actually are.

THEORIES OF REASONING

Evans (1991) offers a four-way classification of reasoning theories and a three-way characterization of the questions they must try to answer. The questions which need to be addressed are: the competence question – the fact that human subjects often successfully solve deductive-reasoning problems; the bias question – the fact that subjects also make many systematic errors; the content-and-context question – the fact that the content and context of a problem can radically alter subjects' responses. Evans (1991) argues that the four theories of reasoning tend to concentrate upon one question or the other, but none provide a fully integrated account of all three. The first two theories address the competence question.

The *mental-logic approach* argues for the existence of formal inference rules in the cognitive system (Inhelder and Piaget, 1958; Henle, 1962; Braine, 1978; Johnson-Laird, 1975; Osherson, 1975; Rips, 1983). These rules, for example, modus ponens, i.e. 'given if p, then q and p you can infer q', rely on the syntactic form of the sentences encoding the premises. Thus, whatever sentences are substituted for p and q the same inferences apply. *Mental-models theory* suggests that the semantic content of the sentences encoding a hypothesis is directly represented in the cognitive system (Johnson-Laird, 1983; Johnson-Laird and Byrne, 1991). It is these contents which are subsequently manipulated in reasoning. Hence the actual meaning of p and q may be important to the reasoning process.

Two further theories are directed at explaining content affects and the errors and biases which infect people's normal reasoning performance. *Pragmatic-reasoning schema theory* proposes inference rules which are specific to particular domains to account for content effects. Cheng and Holyoak (1985), for example, invoke a permission schema to account for the facilitatory effects of thematic content. In these tasks contentful rules about permission relations were employed, for example, 'If you are drinking alcohol, you must be over 18 years of age'. Last, the *heuristic approach* proposes that a variety of systematic errors and biases in human reasoning may be explained by the cognitive system employing a variety of short-cut processing strategies (Evans, 1983a, 1984, 1989).

Evans (1991) was concerned to get reasoning theorists to agree some common ground rules concerning the adequacy of their theories. He does so by providing criteria of theory preference – completeness, coherence, falsifiability and parsimony – by which to judge reasoning theories and seems to view mental models as scoring most highly on these criteria. We will argue that along with these general criteria – common to all scientific domains – limitations on long-term memory retrieval may also provide a valuable criterion by which to assess reasoning theories.

Cognitive limitations have been appealed to in order to account for the biases which occur in people's reasoning. For example, limitations on short-term memory capacity have been appealed to in order to motivate the heuristic approach (Evans, 1983b, 1989) and to explain error profiles in syllogistic reasoning (Johnson-Laird, 1983). Given the prominence of the frame problem in AI, why has it not also been taken as a potential source of constraint on theories of reasoning? We believe there are two reasons. First, no analysis has been provided of these process theories which might indicate that they are profligate with computational resources. Second, when accounting for laboratory tasks the demands of a generalizable theory of inference can be ignored. We now suggest that contemporary reasoning theories are intended to generalize appropriately to other inferential modes.

THE GENERALIZATION ASSUMPTION

Why has the psychology of deductive reasoning been so prominent within cognitive psychology/science? The main reason appears to be the assumption that the principles of human inference discovered in the empirical investigation of explicit inference will generalize to provide accounts of most inferential processes. We call this the *generalization assumption*. The generalization assumption is, for example, implicit in the sub-title to Johnson-Laird's (1983) book *Mental Models: Towards a Cognitive Science of Language, Inference and Consciousness*. Little overt human activity involves deductive inference. Therefore, without the generalization assumption the study of deductive reasoning would warrant little more interest than, say, the psychology of playing Monopoly.

Within artificial-intelligence knowledge representation a similar generalization assumption encountered the problem of *scaling up*. Quite often programs which worked well in *toy domains*, i.e. small well-behaved databases rather like the abstract domains employed in laboratory reasoning tasks, failed when scaled up to deal with larger more realistic databases. This was because the inference regimes in these AI programs were generally computationally intractable but this was only apparent when they were scaled up to deal with more complex, real-world inferential problems. While a prominent issue in AI research (for example, Levesque, 1985, 1988; McDermott, 1986), scaling up has not been an issue in the psychology of reasoning.

Defeasible inference

What is the nature of the inferential processes to which we expect a generalizable theory of inference to generalize? As we have suggested, little overt human activity may involve deduction. However, these overt activities may be supported by implicit inferential processes which are deductive in nature. According to modern cognitivist accounts activities such as text comprehension, classification, categorization, and perception all rely on inferential processes. The inferences which are required in these areas all share a common characteristic: they are *defeasible*. That is, putative conclusions can be *defeated* by subsequent information.

For example, text comprehension relies on implicit inferences from prior world knowledge to elaborate the information given in the text (Bransford and Johnson, 1972, 1973; Bransford *et al.*, 1972; Bransford and McCarrell, 1975; Clark, 1977; Minsky, 1975; Stenning and Oaksford, 1989). These inferences can be defeated by subsequent sentences that contradict earlier conclusions. Theories of concepts designed to capture the family resemblance or prototype structure of human categorization implicitly recognize the defeasibility of semantic knowledge. So, although not all

birds can fly, the prototypical bird is represented as flying, the majority of exemplar birds fly, the probability that a bird flies is high, etc., depending on the theory that one considers (Rosch, 1973, 1975; Medin and Schaffer, 1978; Nosofsky, 1986). Constructivist theories of perception take much of perceptual processing to involve inference to the best explanation about the state of the environment, given perceptual evidence. The possibility of perceptual illusion and error provides evidence for the defeasibility of such inference (Gregory, 1977; Fodor and Pylyshyn, 1981; MacArthur, 1982). Later on we will also see that defeasibility is observed in reasoning experiments (Byrne, 1989; Cummins *et al.*, 1991) and that the ability to account for these phenomena has been appealed to as arguing in favour of a particular theory of reasoning (Johnson-Laird and Byrne, 1991).

At least *prima facie*, the defeasibility of the inferential modes observed in these cognitive domains rules out a deductive approach. It has often been argued that the single most defining characteristic of a deductive system is that a valid inference *cannot* be defeated by subsequent information (e.g. Curry, 1956). That is, deductive validity is *monotonic*. However, many non-standard but equally *logical* accounts of connectives result in *non-monotonic* systems, for example, the Lewis–Stalnaker (Stalnaker, 1968; Lewis, 1973) semantics for the counterfactual conditional provides such a system (Glymour and Thomason, 1984). Hence, just because the inferences to be characterized are defeasible does not of itself exclude a formal, logical approach.

We now introduce a precise analysis of how a particular process theory may transcend the limitations of the cognitive system. This will involve a discussion of computational complexity theory (see, for example, Garey and Johnson, 1979; Horowitz and Sahni, 1978) which provides a characterization of the resources a computational process consumes.

BOUNDED RATIONALITY AND COMPUTATIONAL COMPLEXITY THEORY

How do we know whether or not a process theory transcends the limitations on the cognitive system? For short-term memory capacity, an answer can usually be provided at an intuitive level. Without 'chunking' if a particular process model requires more than 7 ± 2 items to be stored, then short-term memory capacity will be exceeded. However, more implicit cognitive processes, proceeding outside of conscious awareness, are not usually considered to be bounded by short-term memory capacity. How can it be estimated whether a process postulated at this level transcends the abilities of the cognitive system? On the assumption that cognitive processes are computational processes computational complexity theory provides an answer.

Some computational processes are more complex than others requiring

more computational resources in terms of memory capacity and operations performed. There are two approaches to computational complexity: *a priori* analysis and *a posteriori* analysis (Garey and Johnson, 1979; Horowitz and Sahni, 1978). *A posteriori* analysis involves the observation of the run-time performance of an actual implementation of an algorithm, as the size of the input, n, is systematically varied. Such empirical observations can generate approximate values for best-, worst- and typical-case run-times. A more theoretically rigorous approach is to attempt to derive an expression which captures the rate at which the algorithm consumes computational resources, as a function of the size of n. The crucial aspect of this function is what is known in complexity theory as its *order of magnitude*, which reflects the rate at which resource demands increase with n. For present purposes, the relevant resource is the number of times the basic computational operations of the algorithm must be invoked. Orders of magnitude are expressed using the 'O' notation:

$$O(1) < O(\log n) < O(n) < O(n\log n) < O(n^2) < O(n^3) \ldots < O(n^i) \ldots < O(2^n) \ldots$$

For example, O(1) indicates that the number of times the basic operations are executed does not exceed some constant regardless of the length of the input. $O(n^2) < O(n^3) \ldots < O(n^i)$ indicate that the number of times the basic operations are executed is some polynomial function of the input length, such algorithms are *polynomial-time computable* (strictly speaking this class includes all algorithms of order lower than some polynomial function, such as $O(\log n)$, and $O(n\log n)$).

Within complexity theory an important distinction is drawn between polynomial-time computable algorithms ($O(n^i)$ for some n), and algorithms which require *exponential time* (for example, $O(2^n)$ or worse). As n increases, exponential-time algorithms consume vastly greater resources than polynomial-time algorithms. This distinction is usually taken to mark the difference between tractable algorithms (polynomial time) and intractable (exponential time) algorithms. Applying these distinctions to problems, a problem is said to be polynomial-time computable if it can be solved by a polynomial-time algorithm. If all algorithms which solve the problem are exponential time, then the problem itself is labelled 'exponential-time computable'.

An important class of problems whose status is unclear relative to this distinction is the class of *NP-complete problems*. 'NP' stands for *non-deterministic polynomial-time* algorithms. Problems which only possess polynomial-time algorithms that are non-deterministic are said to be 'in NP'. NP-complete problems form a subclass of *NP-hard* problems. A problem is NP-hard if satisfiability reduces to it (Cook, 1971).[2] A problem is NP-complete if it is NP-hard *and* is in NP. There are problems which are NP-hard but are not in NP. For example, the halting problem is undecidable, hence there is no algorithm (of any complexity) which can solve it. However, satisfiability reduces to the halting problem which thus provides an instance of a problem that is NP-hard but not NP-complete. The class of NP-complete problems includes such classic families of problems as the travelling-salesman problems – the prototypical example of which is the task of determining the shortest round-trip that a salesman can take in visiting a number of cities. It is not known whether any NP-complete problem is polynomial-time computable, but it is known that if any NP-complete problem is polynomial-time computable, then they all are (Cook, 1971). All known deterministic algorithms for NP-complete problems are exponential-time, and it is widely believed that no polynomial-time algorithms exist. In practice, the discovery that a problem is NP-complete is taken to rule out the possibility of a real-time tractable implementation. In practical terms this may mean that for some n an algorithm which is NP-complete may not provide an answer in our lifetimes if at all.

Examples

Issues of computational complexity have arisen quite frequently in the history of cognitive psychology and artificial intelligence, perhaps most notably in vision research and risky decision making. Early work on bottom-up object recognition of blocks worlds resulted in the notorious combinatorial explosion (see McArthur (1982) for a review, and Tsotsos (1990) for a more recent discussion of complexity issues in vision research). In research into risky decision making, it was realized very early that complexity issues were relevant. Bayesian inference makes exponentially increasing demands on computational resources even for problems involving very moderate amounts of information. A salutary example is provided by the discussion of an application of Bayesian inference to medical-diagnosis problems involving multiple symptoms in Charniak and McDermott's (1985) introduction to artificial intelligence. Diagnoses involving just two symptoms, together with some reasonable assumptions concerning the numbers of diseases and symptoms a physician may know about, require upwards of 10^9 numbers to be stored in memory. Since typical diagnoses may work on upwards of 30 symptoms, even if every *connection* in the human brain were encoding a digit, its capacity would none the less be exceeded. Such complexity considerations render it highly unlikely that human decision makers are generally employing Bayesian decision theory in their risky decision making. Such results were primarily responsible for the emergence of the heuristics-and-biases approach in the psychology of human decision making (Tversky and Kahneman, 1974).

For our present purposes, the most telling example where complexity

issues have suggested the infeasibility of an approach is in artificial-intelligence knowledge representation (McDermott, 1986). Most AI programs require knowledge to be represented and accessed. Knowledge is represented in logical form and accessing it treated as a logical inference. A problem AI researchers encountered was that world knowledge is invariably *defeasible*. The standard example is 'All birds can fly'. From this rule and the knowledge that 'Tweety is a bird' you may infer that 'Tweety can fly'. However, this rule is defeasible. If you subsequently learn that 'Tweety is an ostrich', then the conclusion that 'Tweety can fly' is defeated. Note that strictly speaking that ostriches can't fly is a *counterexample* to the original generalization. That is, the generalization is false, and hence no valid conclusions can be drawn from it. This may suggest that only exceptionless generalizations should form the contents of world knowledge. However, as we have already indicated, at least at the level of people's common-sense classification of the world, such exceptionless generalizations would not appear to be available to characterize their everyday world knowledge.

The standard approach (e.g. Reiter, 1980, 1985) has been to argue that a closed world assumption should be made. That is, inferences are drawn based on what is in the knowledge base *now*. Informally, when it is learnt that 'Tweety is a bird', as long as a counterexample can not be generated from the current contents of the database, i.e. 'Tweety can not fly' cannot be established, then it is reasonable to infer that 'Tweety can fly'. This means that every time a conclusion is drawn from a default rule the whole of the database must be exhaustively searched to ensure no counterexample is available. This is equivalent to checking the consistency of the database. But consistency checking reduces to the satisfiability problem and is therefore NP-complete. In consequence *an NP-complete problem has to be solved every time a default rule is invoked*. Since in the human case the database may consist of the whole of world knowledge, this logical account looks unpromising.

Of course this is a variant of the frame problem. It would be a great advantage if, rather than exhaustively searching the whole of world knowledge, only some *relevant* subset needed to be checked. The problem is then how to achieve this in a non-arbitrary way. As we will see below, two reasoning theories – pragmatic reasoning schema theory and the heuristic approach – potentially address this problem. However, we will argue that they provide inadequate responses to the problem of intractability.

Summary

Let us sum up the argument so far. We have suggested that considerations of bounded rationality may serve to provide criteria by which to judge

current theories of reasoning. The reason why such considerations have not been taken into account is a failure to address the generalization issue. That is, theories of laboratory tasks must be able to generalize to more realistic inferential contexts. This is analogous to the problem of 'scaling up' in AI knowledge representation: many inference theories are suitable only to 'toy', or alternatively, 'un-ecologically' valid, domains. The majority of real human inference is defeasible or non-monotonic. However, standard approaches to defeasible inference would appear to be computationally intractable because of their reliance on exhaustive searches for counterexamples. In the following section, we will discuss the four theories of reasoning introduced above in the light of these considerations. As we said above, we will argue that all these theories of reasoning either make unreasonable demands on cognitive resources or provide inadequate responses to the problem of cognitive limitations.

THEORIES OF REASONING AND BOUNDED RATIONALITY

We will deal with the four theories of reasoning in the order they were introduced: mental logics, mental models, pragmatic reasoning schemas, and the heuristic approach.

Mental logics

The contemporary mental-logic view explains explicit reasoning performance by appeal to various natural deduction systems (Gentzen, 1934) with (Braine, 1978), or without (Rips, 1983) some specific assumptions concerning the processes which animate the inference rules.[3] From the perspective of computational complexity, mental-logic accounts appear particularly unpromising. Even for standard monotonic logics, the general problem of deciding whether a given finite set of premises logically implies a particular conclusion is NP-complete (Cook, 1971).[4] Moreover, the *a priori* complexity results discussed above were derived from logical attempts to account for default reasoning in AI knowledge representation. In consequence, it seems unlikely that the mental-logic approach is going to satisfy the generalization assumption. There would appear to be only two possible lines of retreat to avoid the conclusion that most inferential performance is beyond the scope of the mental-logic approach.

First, despite *a priori* arguments that most human reasoning is defeasible, people may employ a standard logic in much everyday reasoning. However, over the last 30 years or so it has been the failure to observe reasoning performance that accords well with standard, monotonic logic which has led to questions over human rationality. When as little as 4 per cent of subjects' behaviour accords with standard logic in tasks where it is appropriate, it seems odd to generalize such an account

to situations where it is not. Nevertheless, it must be conceded that this is an empirical issue. People *may* treat everyday defeasible claims as exceptionless generalizations. This possibility is, however, sufficiently remote for us to consider it no further.

Second, the generality of mental logics may be restricted to explicit reasoning and it may be denied that they are intended to cover implicit inferential processes involved in common-sense reasoning. Intractability is therefore not an issue because of the small premise sets involved. This proposal of course explicitly denies that mental logics can satisfy the generalization assumption. It, moreover, may not save the mental-logic account from intractability problems. Above we suggested that it is highly unlikely that standard monotonic inference is generalized to everyday defeasible inference. We now argue that the converse is far more plausible, i.e. that explicit reasoning may be influenced by defeasible inferential processes. If this is the case then explanations of human inferential behaviour, even on explicit reasoning tasks, will have to address the tractability problems we have raised.

The proposal that explicit reasoning may be influenced by defeasible inferential processes derives from recent empirical work on conditional reasoning. It would appear that even in laboratory tasks conditional sentences may be interpreted as default rules (Oaksford *et al.*, 1990). Byrne (1989) and Cummins *et al.* (1991) have shown that background information derived from stored world knowledge can affect inferential performance (see also Markovits, 1984, 1985). Specifically they have shown that the inferences which are permitted by a conditional statement are influenced by *additional antecedents*. For example:

1 If the key is turned the car starts.
  (a) Additional antecedent: the points are welded.

(1) could be used to predict that the car will start if the key is turned. This is an inference by modus ponens. However, this inference can be *defeated* when information about an additional antecedent (a) is explicitly provided (Byrne, 1989). Moreover, confidence in this inference is reduced for rules which possess many alternative antecedents even when this information is left implicit (Cummins *et al.*, 1991). In these studies additional antecedents were also found to affect inferences by modus tollens. If the car does not start, it could be inferred that the key was not turned, unless, of course, the points were welded. Modus tollens is *defeated* when information about an alternative antecedent is explicitly provided (Byrne, 1989) and confidence in it is reduced for rules which possess many alternative antecedents even when this information is left implicit (Cummins *et al.*, 1991).

The rules employed in these laboratory tasks are being treated as default rules. Other evidence indicates that even abstract rules may be treated

in this way. In conditional inference tasks (Taplin, 1971; Taplin and Staudenmayer, 1973) and Wason's (1966) selection-task subjects typically refrain from either drawing inferences that accord with modus tollens or adopting the strategy of falsification that is sanctioned by modus tollens. This can be at least partially explained if it were a general default assumption that all rules are default rules. If this were the case, then modus tollens may be suppressed because the rules are treated as defeasible, just as in Byrne (1989) and Cummins *et al.* (1991).[5]

In sum, it seems likely that conditionals employed in explicit reasoning tasks are treated as default rules. Restricting the applicability of mental-logic approaches to explicit reasoning does not, therefore, avoid the problems of computational intractability.

The influence of default rules on people's reasoning would appear to have been dismissed by mental logicians as interfering pragmatic or performance factors (Rumain *et al.*, 1983; Braine *et al.*, 1984). This is in marked contrast to the reaction of logicians and AI researchers. These researchers have almost uniformly abandoned restrictions on what is deducible to the monotonic case and have been exploring non-monotonic logics to capture just the phenomenon their mental counterparts dismiss (see, for example, the collection edited by Ginsberg, 1987). The intuition behind this reaction seems to be that unless logical methods can be applied to these cases then most interesting inferences may be beyond the scope of logical inquiry. Logical enquiry may proceed divorced from the requirement to provide computationally tractable inference regimes. Most AI applications and the cognitive science of human reasoning cannot, however, avoid these problems.

In conclusion, providing a viable theory of human inference must resolve the issue of intractability. Unfortunately a solution does not appear to be forthcoming from within the formal, logical approach. This is not incompatible with continued logical enquiry into systems which can handle default reasoning. Further, the possibility can not be dismissed that some formal notation may be devised which allows for more tractable implementations. However, the lack of practical success in devising a tractable logic for default inference suggests that this may be what Lakatos (1970) referred to as a degenerative research programme (Oaksford and Chater, 1991). In consequence, it seems unlikely that the mental-logic approach will satisfy the generalization assumption.

Mental models

The apparent failure of logical accounts to generalize appropriately to everyday common-sense inference appears to add further weight to the mental modeller's claim that 'there is no mental logic'. On the mental-models view, the syntactic formalisms adopted by the mental logician

should be abandoned in favour of semantic methods of proof (e.g. Johnson-Laird, 1983; Johnson-Laird and Byrne, 1991). Such methods do not possess formal, syntactic rules of inference like modus ponens or modus tollens. Rather, the semantic contents of premises are directly manipulated in order to assess whether they validly imply a conclusion.

In this section we will introduce two interpretations of mental models. One we refer to as 'logical mental models', the other as 'memory-based mental models'.

Logical mental models

In recent accounts of mental models the claim that 'there is no mental logic' has been tempered. For example, 'the [mental] model theory is in no way incompatible with logic: it merely gives up the formal approach (rules of inference) for a semantic approach (search for counterexamples)' (Johnson-Laird and Byrne, 1991: 212). So the dispute is not about *whether* there is a mental logic, but about *how* it is implemented. On this interpretation *logical* mental models may be seen as an attempt to provide the notation, to which we alluded above, that will allow a tractable implementation of logic.

Mental models contrast with some semantic approaches to searching for counterexamples but share similarities with others. Truth tables and semantic tableaux (e.g. Hodges, 1975), which are unquestionably logical,[6] contrast with mental models because they are defined over standard propositional representations. In this respect mental models are more related to graphical proof methods such as Euler circles and Venn diagrams. In these semantic-proof procedures the operations which correspond to the steps of a sound logical derivation are defined over graphical representations of the domains of the quantifiers.

As Evans (1991) observes, both the mental-logic approach and mental models are attempting to account for human deductive competence. In assessing the mental-models approach, it would be helpful, therefore, if answers could be found to the same *metatheoretical* questions concerning computational tractability that we asked of the mental-logic approach. Certainly on the *logical* mental-models interpretation, answers to these questions should be possible. However, none as yet would appear to be available. This makes it difficult to assess mental models by the same standards we have applied to mental logics. This is a general problem. While mental models are supposed to do the same job as a mental logic, there are no metatheoretical proofs that this is the case. None the less, in the absence of the appropriate proofs, we can speculate about how the answers to these questions may turn out.

The first tractability question we looked at with mental logics was the standard case of monotonic inference where we found that the general

problem of deciding validity was NP-complete. While this is generally the case, the situation is even worse with standard 'semantic approache[es]'. At this point we must head off a possible confusion. The semantic methods we mentioned above – truth tables and semantic tableaux – are formal *proof* methods (Hintikka, 1985). In contrast, the intention behind the 'semantic approach' of mental models is to use *model theory* as a basis for inference. As Hintikka (1985) observes, model theory *per se* provides no inferential mechanisms. However, the models could be exhaustively checked. For example, the sentence 'Gordon is in his room' (indexed to a particular space-time location, say *now*) will be true if and only if Gordon is in his room now, i.e. Gordon actually being in his room now provides *a* model for this sentence. Of course, this is a contingent claim and therefore there are many models in which it is false. Nevertheless you could check this sentence is true by looking at the arrangement of objects about which the claim is made. Could you check the validity of a putative logical truth in a similar way? Logical validity is defined relative to *all* models, which are potentially infinite in number. Moreover, many of them will be infinite in size. Attempting to prove the logical validity of a statement in this way would be impossible, at least for the finite minds of human beings. In sum, basing a psychological theory of inference on model theory looks even less promising than using formal syntactic methods.

Mental-models theorists are well aware of this problem (Johnson-Laird, 1983) and argue explicitly that mental models may provide a way in which model theory may be developed in to a tractable proof procedure. Mental models only deal with small sets of objects which represent *arbitrary exemplars* of the domains described in the premises. This is analogous to Bishop Berkeley's claim that reasoning regarding, say triangles, proceeds with an arbitrary exemplar of a triangle, rather than the, in his view, obscure Lockean notion of an abstract general idea. Providing no assumptions are introduced which depend on the properties of this particular triangle, for example, that it is scalene rather than equilateral, then general conclusions concerning *all* triangles may be arrived at.

The introduction of arbitrary exemplars highlights the lack of an appropriate metatheory for mental models. There is no exposition of the rules which guarantees that no illegitimate assumptions are introduced in a proof. This does not mean that any particular derivation using mental models has made such assumptions. None the less, guaranteeing the validity of an argument depends on ensuring that in a particular derivation one *could* not make such assumptions. Hence explicit procedures to prevent this happening need to be provided. In their absence there is no guarantee (i.e. no proof) that the procedures for manipulating mental models preserve validity. That is, it is not known whether, relative to the standard interpretation of predicate logic, mental models provides a *sound* logical system.[7]

While soundness is unresolved, there are strong reasons to suppose that mental models theory is not *complete* with respect to standard logic, i.e. while all inferences licensed by mental models may be licensed by standard logic (soundness) the converse is not the case. Other *graphical* methods are restricted in their *expressiveness* due to physical limitations on the notation. Venn diagrams, for example, can only be used to represent arguments employing four or less *monadic* predicates, i.e. predicates of only one variable (Quine, 1959).[8] They therefore only capture a small subset of logic. While mental models have been used to represent relations, i.e. predicates of more than one variable, there is no reason to suppose that mental models will not be subject to analogous limitations. If so, then mental models will not provide a general implementation of logic.[9]

The employment of arbitrary exemplars is central to providing a tractable model-based proof procedure. However, there are no complexity results for the algorithms which manipulate mental models. Such demonstrations may be felt unnecessary, if, as with the mental-logic approach, mental-models theory were restricted to the explicit inferences involved in laboratory tasks. However, mental-models theory has been generalized to other inferential modes, including implicit inference in text comprehension (Johnson-Laird, 1983). As we mentioned above, these inferences are defeasible (see p. 34), as are most everyday inferences people make.[10] Further, in many laboratory reasoning tasks, conditional sentences would appear to be interpreted as default rules (see above). So in order to provide a general theory of inference, mental models must account for defeasibility.

Proposals for incorporating default reasoning into mental models (Johnson-Laird and Byrne, 1991) rely on incorporating default assumption into the initial mental model of a set of premises. These assumptions will be recruited from prior world knowledge and may be undone in the process of changing mental models. The problem of consistency checking can be avoided because no search for counterexamples to these default assumptions need be initiated. This proposal does not resolve the problem of default inference. A generalizable theory of reasoning must address the problem of *which* default assumption(s) to incorporate in an initial representation. For example, suppose you are told 'Tweety is a bird', you may incorporate the default assumption that 'Tweety can fly' in your mental model because most birds can fly. However, it would be perverse to incorporate this assumption if you also knew that 'Tweety is an ostrich'. To rule out perverse or *irrelevant* default assumptions requires checking the whole of world knowledge to ensure that any default assumption is consistent with what you already know (or some relevant subset of what you already know). This will involve an exhaustive search over the whole of world knowledge for a counterexample to a default assumption.

It could be argued that the problem of searching for counterexamples

for default assumptions is part of the theory of memory retrieval which mental models, as a theory of inference, is not obliged to provide. Three arguments seem to vitiate this suggestion. First, as we have seen, in AI at least, these memory-retrieval processes are treated as *inferential* processes and therefore need to be explained by a theory of inference. Second, the memory-retrieval processes involve the search for counterexamples. This indicates that *in its own terms* they are exactly the kind of inferential processes for which mental-models theory should provide an account. Third, such an argument could only succeed if mental-models theory itself didn't already rely heavily on such processes to explain the results of reasoning tasks.

In recent accounts (e.g. Johnson-Laird and Byrne, 1991) the explanation of various phenomena depend on the way in which an initial mental model of the premises is 'fleshed-out'. Fleshing-out, for example, determines: (i) whether a disjunction is interpreted as exclusive or inclusive, or (Johnson-Laird and Byrne, 1991: 45) (ii) whether a conditional is interpreted as material implication or equivalence (Johnson-Laird and Byrne, 1991: 48–50), which in turn determines whether inferences by modus tollens will be performed; (iii) whether non-standard interpretations of the conditional are adopted (Johnson-Laird and Byrne, 1991: 67), including content effects whereby the relation between antecedent and consequent affects the interpretation (Johnson-Laird and Byrne, 1991: 72–3); (iv) confirmation bias in Wason's selection task (Johnson-Laird and Byrne, 1991; 80) and (v) the search for counterexamples in syllogistic reasoning (Johnson-Laird and Byrne, 1991: 119). Fleshing out depends on accessing world knowledge. Moreover, the explanatory burden placed on fleshing out demands that mental-models theory accounts for the processes involved. In consequence it is reasonable to expect mental-models theory to provide an account of how relevant defaults are also retrieved from world knowledge. Since this issue is not addressed it seems unlikely that logical mental models can satisfy the generalization assumption.

However, the processes of fleshing out may suggest another interpretation of mental models which we briefly present before closing this section.

Memory-based mental models

The explanatory burden placed on fleshing out suggests that the memory-retrieval processes involved may be primarily responsible for mental-model construction and manipulation. The representations that appear in, for example, Johnson-Laird and Byrne (1991) may be better regarded as the *products* of processes in which those representations are not explicitly involved. In other words they are the 'appearance(s) before the footlights of consciousness' (James, 1950/1890) of processes which are not

defined over those representations themselves. This contrasts with logical mental models where the processes that transform one model into another *are* defined over the representations that appear on the pages of, for example, Johnson-Laird and Byrne (1991).

Memory-based mental models appear to accord with an earlier thread in mental models theory:

> Like most everyday problems that call for reasoning, the explicit premises leave most of the relevant information unstated. Indeed, *the real business of reasoning in these cases is to determine the relevant factors and possibilities*, and it therefore depends on knowledge of the specific domain. Hence the construction of putative counterexamples calls for an active exercise of memory and interpretation rather than formal derivation of one expression from others.
>
> (Johnson-Laird, 1986: 45, our emphasis)

On a memory-based mental-models position the 'active exercise of memory and interpretation' would represent the heart of all inferential processes. Moreover, existing accounts of mental models could be interpreted as specifying the intended outputs of these processes given certain inputs. In this respect mental-models theory could therefore be expected to provide a valuable source of constraint on a future memory-based theory of reasoning. We will return to this interpretation of mental models later on.

Summary

Recent accounts of mental-models theory appear to favour an interpretation in terms of a graphical semantic-proof procedure. On this interpretation, mental models provides an alternative notation for implementing logic in the mind. This invites a variety of *metatheoretic* questions which need to be answered to assess the adequacy of *logical* mental models as a general, tractable, implementation of logic. Unfortunately, answers to these questions are unavailable. Further, existing proposals for handling default inference are inadequate. Taken together these considerations argue for a Scots verdict of 'not proven' on logical mental models. However, the processes of fleshing out indicate that memory-based mental models, while less articulated, may act as a valuable source of constraint on a memory-based theory of inference.

Pragmatic-reasoning schema theory

Pragmatic-reasoning schema theory emphasizes the role of domain-specific knowledge in reasoning tasks (Cheng and Holyoak, 1985; Cosmides, 1989). Cheng and Holyoak (1985) suggested that people possess *pragmatic reasoning schemas*, which embody rules specific to various domains such as permissions, causation, and so on. Permission schema are invoked in explaining the results from some thematic versions of Wason's selection task where the rule determines whether or not some action may be taken. Cheng and Holyoak (1985) argue that the rules embodied in a permission schema match the inferences licensed by standard logic, thus explaining the facilitatory effect of these materials. Similarly, Cosmides (1989) appeals to domain-specific knowledge of 'social contracts' to explain the same data (but see Cheng and Holyoak, 1989, for a critique). While Cosmides' work on social contracts is important, it is only the postulation of data structures specific to particular domains which will concern us.

We have frequently remarked that if the domains over which the search for counterexamples takes place were suitably constrained, then exhaustive searches may be feasible. However, there are two reasons for suspecting that schema-theoretic or domain-specific approaches in general will not prove adequate.

First, default reasoning is about how beliefs are appropriately updated in response to new information (Harman, 1986). Within philosophy the processes involved have typically been discussed under the heading of confirmation theory (Fodor, 1983). In arguing that confirmation, and hence default reasoning, is subject to the frame problem, Fodor observes that confirmation is characteristically *isotropic*:

> By saying that confirmation is isotropic, I mean that the facts relevant to the confirmation of a scientific hypothesis may be drawn from anywhere in the field of previously established empirical (or, of course, demonstrative) truths. Crudely: everything that the scientist knows is, in principle, relevant to determining what else he ought to believe.
>
> (Fodor, 1983: 105)

Domain specificity can assist with intractability only if isotropy is abandoned. If default reasoning is isotropic, then placing rigorous boundaries on relevant information would be a move in exactly the wrong direction. A knowledge organization which excluded the possibility of isotropy would be hopelessly inflexible. Although cross-referencing schemata is a possibility, as Fodor (1983: 117) points out: 'an issue in the logic of confirmation . . . [becomes] . . . an issue in the theory of executive control (a change which there is, by the way, no reason to assume is for the better)'.

A second reason to suspect that domain-specific approaches are inadequate concerns the lack of any general principles concerning how an appropriate compartmentalization of knowledge is to be achieved. Such general principles are required since otherwise how knowledge is organized into discrete compartments from the flux of information that an organism receives in interacting with its environment remains opaque (Oaksford

and Chater, 1991). While it may be legitimate to appeal to compartmentalization, once appealed to, an account of how it is achieved must be supplied. Pragmatic-reasoning schema theory does not explicitly address this issue. In consequence it is unlikely that this theory can satisfy the generalization assumption.

Heuristic approaches

The heuristic approach (Evans, 1983b, 1984, 1989) is that most concerned with the issue of cognitive limitations (Evans, 1983a). In computer science the use of heuristics may render a computationally intractable problem manageable. Tractable, approximate solutions may be found for many problem instances by employing the generally intractable algorithm with a heuristic (Horowitz and Sahni, 1978). Accuracy is traded for speed. In this section we will observe that the current heuristic approach does not address the intractability problems we have raised: the heuristics proposed are more often motivated by appeal to *pragmatic* rather than *processing* factors. We will suggest, however, that with some minor reinterpretation, one heuristic proposed by Evans (1983b) may address the intractability issue. None the less, we will conclude that supplementing generally intractable algorithms with heuristics is unlikely to provide a general solution to the problem of intractability.

The *not*-heuristic (Evans, 1983b, 1984, 1989) is motivated by Wason's (1965) proposal that negations are typically used to deny presuppositions. For example, 'I did *not* go for a walk' denies the presupposition that you went for a walk. The topic of this sentence – what the sentence is about – is walking and not any of the things I could have done while not walking. On the basis of this example it was proposed that the language understanding mechanism embodies a *not*-heuristic (Evans, 1983b). This heuristic treats information about, for example, what you did while *not* walking as irrelevant. Attention is therefore focused only on the named values. More recently, this heuristic has been regarded as a manifestation of a general bias towards positive information, i.e. information about what something is rather than what it is not (Evans, 1989; see also Oaksford and Stenning, 1992).

Such a general preference for positive information may be better motivated by processing rather than pragmatic considerations. A general positivity bias may be one aspect of providing a tractable knowledge base (Oaksford and Chater, forthcoming). The frame problem was first noticed in reasoning about change. In a dynamic representation, the consequences of something changing has to include all the things that did *not* change. For example, along with the information that 'If your coffee cup is knocked over your carpet gets wet', all the information about what did not happen when your coffee cup is knocked over needed to be encoded. For example,

that the window does not open, the lights do not switch off and so on. There is a potentially infinite list of things which do not happen as a consequence of knocking your coffee cup to the floor, each of which would have to be explicitly represented. However, the *negation-as-failure* procedure obviates the need to represent all this information (Hogger, 1984).[11] If, from the current contents of the database, it cannot be proved that the window opens, then it is assumed that the window does not open. The upshot is that in a logic program *no* negative information is stored (Hogger, 1984). This represents a prime case of positivity bias in the service of tractability.

So at least one aspect of the current heuristic approach could address the tractability issues we have discussed. However, as Evans (1991) says, the heuristic approach is *not* an approach to human reasoning in its own right. It needs to be married to a particular theory of competence. Such an approach is unlikely to prove adequate, however. The problem is that:

> The use of heuristics in an existing algorithm may enable it to quickly solve a large instance of a problem provided the heuristic 'works' on that instance. . . . A heuristic, however, does not 'work' equally effectively on all problem instances. Exponential time algorithms, even coupled with heuristics will still show exponential behaviour on some sets of inputs.
>
> (Horowitz and Sahni, 1978)

There has been no attempt to articulate the sets of heuristics which would be needed to provide generally tractable inference regimes either within the heuristic approach or in AI knowledge representation. Hence, Evans (1991) may well be right that one way to proceed is to marry the heuristic approach to one or other of the theories which explicitly address the competence issue. However, it seems doubtful that an appropriate set of heuristics will be forthcoming to supplement these theories (Oaksford and Chater, 1991).

Default reasoning in particular presents new problems for the heuristic approach. Existing accounts of default reasoning fail to arrive at intuitively acceptable conclusions (McDermott, 1986). Quite often the only conclusion available is of the form p v not-p, i.e. a logical truth (Oaksford and Chater, 1991). This is particularly uninformative. It has been suggested that one way to resolve this problem is by appeal to various heuristics. These heuristics may also assist with tractability by cutting down the number of possibilities which need to be considered. The disjunction above is all that can often be concluded because each default rule may lead to a different possible conclusion. Logically, the only conclusion that can be drawn therefore is their disjunction. However, if one default rule can be given preference, then all these possibilities need not be computed (see Oaksford and Chater, 1991).[12] Again, however, it is not at all clear that

any of the heuristics proposed resolve this issue appropriately for all instances of a problem (Loui, 1987). In sum, it seems unlikely that an appropriate set of heuristics will be forthcoming to solve the problem of computational intractability. In consequence, the heuristic approach is unlikely to satisfy the generalization assumption.

Summary

In this section we have surveyed existing theories of reasoning with respect to their ability to generalize appropriately to everyday common-sense reasoning. The mental-logic approach was perhaps the least promising in this respect. This is largely because it is sufficiently well articulated for the relevant metatheoretic results to be available. This was in contrast to the logical mental-models approach. Although there is a possibility that arbitrary exemplars may provide for a tractable model-based inference regime, the absence of the relevant metatheoretic results means that it is impossible to decide one way or the other. However, when it comes to default reasoning the mental-models approach is demonstrably inadequate: the real problem is avoided. The possibility remains that memory-based mental models may none the less be explained as emergent properties of a theory of memory retrieval (this possibility is discussed further below). The two theories perhaps most suited to addressing the tractability issue – pragmatic-reasoning schema theory and the heuristic approach – were equally unpromising. Without an account of how compartmentalization is achieved, schema theoretic approaches *presuppose* a solution, they do not provide one. It moreover seems unlikely that an appropriate set of heuristics can be specified to resolve the intractability problem.

DISCUSSION

There are two broad areas which require further discussion in the light of the above arguments. Both concern the issue of rationality. First, we will discuss philosophical implications for human rationality. Second, we will discuss the implications for psychological theories concerned to build rational mechanisms (Fodor, 1987).

Rationality

In this section we will discuss two issues, the implications of reasoning data for human rationality, and the possible charge that abandoning rule-based theories leads to relativism.

The intractability results we have reported indicate that a bounded-rationality assumption should be made. This has the consequence that the empirically observed deviations from normative theories could not bring human rationality into question. The complexity results we have discussed indicate that people *could not* generally be using the normative strategy. It is only possible to condemn people as irrational for not using a particular strategy if they *could* use it. To think otherwise, would be like condemning us because we can not breathe under water even though we do not possess gills. It could be argued, however, that for laboratory tasks involving just a few premises complexity issues are not a concern. We have partly replied to this response above where we observed that if just one rule is interpreted as a default rule, a feasible real-time inference is doubtful. It also seems highly unlikely that people have been endowed with all the logical machinery spontaneously to solve just those tasks small enough not to tax their limited resources. If nothing else this is because the empirical data appear to indicate that they just happen not to use that machinery! It seems far more parsimonious to suggest that the strategy which is used in everyday reasoning contexts is generalized to laboratory tasks.

It would be irrational to demand that people employ strategies which they are incapable of using. However, one attractive feature of rule-based theories is that they come with their own warrant of rationality, as it were. Brown argues that '[on] our classical conception of rationality . . . the rationality of any conclusion is determined by whether it conforms to the appropriate rules' (Brown, 1988: 17). If rule-based theories are abandoned, there may be no guarantee that the strategies which replace them are rational: since they will not be rule-based, they will not carry their own warrant of rationality. This, moreover, may be seen as the first step on the slippery slope towards *relativism*, i.e. the view that there are no universal principles of rationality.

Johnson-Laird and Byrne (1991) consider the same problem and conclude that rather than conformity to rules, the search for counter-examples provides a universal principle of rationality. However, this provides neither a necessary nor a sufficient condition for rational judgement. It is not necessary because it is not a principle universally adhered to in scientific practice which provides our paradigm case of rational activity (Brown, 1988). Within periods of normal science (Kuhn, 1962) scientists explicitly refuse to allow core theoretical principles to be subject to refutation. The search for counterexamples is also not a sufficient criterion for rational judgement. Continuing to search for counterexamples indefinitely is not rational when trying to reach a decision in real time.

However, the idea that the search for counterexamples provides a universal criterion of rationality need not be wholly abandoned. It will, however, need to be supplemented by a theory of *judgement*: 'Judgement is the ability to evaluate a situation, assess evidence, and come to a reasonable decision without following rules' (Brown, 1988: 137). It is a matter of judgement, for example, when and if counterexamples are

allowed to falsify a core theoretical principle, or when the search for counterexamples has been sufficiently exhaustive. Quite frequently we appeal to experts, who have a wealth of experience and knowledge in order to make these judgements. A good example is the peer review system. There is no algorithm for determining whether an experimenter has made sufficient attempts to dismiss alternative explanations of a hypothesis. In consequence, it is left to a researcher's peers to decide whether she/he has adequately dealt with the *relevant* possibilities. A further example is provided by the legal concept of *precedent*. In certain cases a defence lawyer will seek to find a case in which the facts are as similar as possible and where a not-guilty verdict was returned. Equally, the prosecution may seek a similar case where a guilty verdict was returned. Both defence and prosecution are searching for counterexamples to each other's arguments that on the basis of the evidence the defendant should (or should not) be convicted. Judgement enters in to the decision process, in two ways. First, the judge of the present case must decide whether the cases are similar in the *relevant* respects. Second, the whole concept of precedent relies on allowing previous judgements to influence subsequent judgements.

In sum, the claim that we could not employ rule-based theories could lead to relativism. The search for counterexamples *per se* is an inadequate response to this charge. The examples we adduced indicate that the search for counterexamples must be supplemented by a theory of judgement before anything like a universal principle is available.

Rational mechanisms

Rule-based systems operating over formal symbolic representations have the advantage that they possess a transparent semantics which allows us to see how mental representations can be causally efficacious in virtue of their meaning (Fodor, 1987). If we abandon rule-based theories, do we also abandon the ability to provide causal, mechanistic explanations of the way representational mental states mediate behaviour? Part of an answer to this question has already been provided. If the concept of what it is to be rational changes, then the form that a theory of rational mechanism must take may also change. We now consider what kinds of mechanism may be consistent with our developing conception of rationality. We will first draw on an analogy with Kahneman and Tversky's work on risky decision making, and then propose that connectionist systems may provide alternative rational mechanisms.

In response to similar complexity results for Bayesian inference, Tversky and Kahneman (1974) proposed a qualitatively different theory to explain risky decison making in which the normative theory was not retained in any form. The problem of deriving probability estimates was radically reconceived largely in terms of the processes of memory retrieval. Their

heuristic approach can be contrasted with the heuristic approach in theories of reasoning. As we mentioned above, within reasoning theory, heuristics are regarded as supplements to a theory of competence (Evans, 1991). However, in Kahneman and Tversky's approach various memory-based heuristics are regarded as wholesale replacements for the competence theory. We suggest that confronted with similar intractability problems reasoning theorists should adopt the same response.

What could represent an analogous reconceptualization of reasoning mechanisms? Levesque (1988) has suggested that connectionism may represent one strategy in the attempt to develop plausible cognitive mechanisms for inference. Rumelhart *et al.* (1986) and Rumelhart (1989) have also suggested that a predictive neural network may form the basis of people's reasoning abilities. What kind of reconceptualization of reasoning does this involve?

Inference is the dynamics of cognition. In classical approaches (Fodor and Pylyshyn, 1988; Chater and Oaksford, 1990) inference takes static symbolic representations and turns them to useful work, predicting the environment, explaining an experiment, drawing up a plan of action and so on. Formal inference over language-like representations has seemed the only way in which meaning and mechanism could combine (Fodor, 1987). Connectionism may offer a very different picture of how to achieve the marriage between mechanism and meaning. Logic provides a dynamics for representations of a particular type: atomic symbolic representations usually map one to one onto our common-sense classification of the world. Connectionism postulates distributed representations of a very different kind in which stable patterns of features represent items in that classification. The dynamics of the system, moreover, is defined at the featural level and owes more to statistical mechanics than to logic. Nevertheless it may be that these representations and the dynamics which transforms one such representation into another can form the basis of a theory of inference.

Let us consider the problem at a higher level of abstraction. Inference leads us from one interpreted mental state to another. The heart of the problem is how to get mental states to systematically track states of the world or, in other words, how to get the dynamics of cognition to 'hook up' to the dynamics of the world (Churchland and Churchland, 1983). We see no reason, *a priori*, why connectionist systems cannot also perform this function.

While there are serious problems for a connectionist theory of inference, there may also be advantages. It may be compatible, for example, with the second interpretation of mental models we offered above (Rumelhart, 1989). Given a set of inputs a network settles on an interpretation which least violates the constraints embodied in its weighted connections between units. These weighted connections embody the network's knowledge of a domain. One way of characterizing such a relaxation search, is that prior

to input clamping all the knowledge that is embodied in the network is potentially relevant to interpreting the input. However, as the net relaxes into an interpretation only those items most relevant will remain on. The stable state arrived at can be regarded as the initial 'mental model' of the input. This model may embody default assumptions. For example, in the 'on-line' schema model (Rumelhart *et al.*, 1986), a constraint satisfaction network embodied information about prototypical rooms. If the bath unit was clamped on then units like toilet, toothbrush, and so on would come on as default values. In the search for counterexamples, intermediate mental models may be generated by selectively clamping off units and allowing the net to settle into a new stable state (Rumelhart, 1989).

Further, this mode of operation seems to capture something of what it means to make a *judgement*. As we said above, determining whether relevant counterexamples have been exhausted is a matter of judgement based upon what you know. In a simple connectionist system all that it knows (all its synaptic weights) contribute to determining what is relevant to interpreting current inputs. The example of precedent also indicates that counterexamples to *novel* situations may be sought by reference to *similar* situations. The partial pattern-matching capabilities of networks make them good candidates for implementing the processes responsible.[13]

The burden of complexity may also be located in the right place. Within connectionist systems learning is the computationally expensive process. Once learnt, however, an inference over the representations embodied in the network is effortless. In contrast, in classical systems inference is computationally expensive while learning is an issue rarely addressed. This may seem like just trading one complexity problem for another. However, the connectionist system at least mirrors the difficulty people actually appear to encounter with learning and inference.

There are serious problems, however. Current network dynamics are insufficiently articulated to provide an account of the productivity of language and thinking (Fodor and Pylyshyn, 1988). In particular, thinking is not a purely predictive process which is triggered by external events. Indeed in thinking people appear able to 'un-hook' the dynamics of cognition from the dynamics of the world, enabling them to step out of real time. This will require networks to have their own intrinsic dynamics to allow thoughts to chain together in the absence of provoking stimuli. While posing a serious problem there is, none the less, a great deal of work going on in this area (Chater, 1989; Elman, 1988; Jordan, 1986; Rohwer, 1990). We see no reason to be pessimistic about its outcome and the consequent prospects for a connectionist theory of inference.

CONCLUSIONS

We have argued that an adequate theory of reasoning must be able to 'scale up' to deal with everyday defeasible inferences in real time. We observed that no contemporary theory of reasoning provided a tractable account of everyday inference and that in consequence none of these theories were likely to be psychologically real. Concentration on limited laboratory tasks would appear to have led to the development of theories of dubious ecological validity. Further, it would appear more likely that people 'scale down' their everyday strategies to deal with laboratory tasks and that this is the source of the systematic biases observed in human reasoning. While these arguments do not bring human rationality into question, they do demand a reconceptualization of appropriate mechanisms for inference. We suggested that connectionist systems may be appropriate which appeared consistent with memory-based mental models and the requirements of a theory of judgement.

In conclusion, empirical research into human reasoning may need to be more ecologically valid. The boundaries of *real inference* need to be mapped out: how do people deal with defeasible knowledge, how do they make relevance judgements, and how does background information (Byrne, 1989; Cummins *et al.*, 1991) interact with reasoning processes? Answers to these questions could be pursued on two fronts. First the complexification of the laboratory situation. Most reasoning tasks are still pencil-and-paper exercises (although, see Mynatt *et al.*, 1977, for example). In contrast the computer game may offer the prospect of engaging subjects in novel dynamic environments over which the experimenter has control. In such environments, context-sensitive rules, varying difficulties of obtaining information, and differing utilities for correct inference can be arranged and their consequences for behaviour mapped out. Second, more direct analyses of real inferential settings such as the court room and science itself need to be conducted (e.g. Tukey, 1986; Tweney, 1985). Explaining the inferential processes that obtain in such real-world settings must be the ultimate goal of a psychological theory of reasoning.

ACKNOWLEDGEMENTS

We gratefully acknowledge the support of the Economic and Social Research Council, U.K., Contract No. R000231282, in conducting the research which led to this paper.

NOTES

1 It is important to be clear about whose inferential behaviour reasoning theorists are attempting to explain. Throughout this chapter it is assumed to be

the spontaneous, unassisted, inferential performance of logically untutored subjects. By 'spontaneous and unassisted' we mean that the subjects are not allowed to use aids such as pencil, paper or computer to make calculations nor are they able to consult with friends or experts. By 'logically untutored' we mean that subjects should have no explicit formal logical training. In other words reasoning theorists are attempting to explain the reasoning abilities which people possess solely in virtue of genetic endowment and general education.

2 The satisfiability problem is to determine whether a formula is true for some assignment of truth values to the variables. 'Reduces' is a technical term of complexity theory (see Horowitz and Sahni, 1978: 511).

3 Natural deduction systems contain no axioms and all inferences are drawn by the application of various inference-rule schemata, e.g. p OR q, not-p $\models$ q (where '$\models$' can be informally glossed as 'therefore').

4 This applies equally well to semantic-proof procedures, such as truth tables and semantic tableaux, as to syntactic procedures such as axioms or natural deduction systems.

5 This would appear to predict that inferences by modus ponens should also be suppressed in these tasks, which is not the case. We examine this issue in more detail elsewhere (Oaksford and Chater, forthcoming).

6 We should also note that under standard interpretations, the search for counterexamples does not distinguish syntactic from semantic approaches. All proof procedures are regarded as 'abortive counter-model constructions' (Beth, 1955; Hintikka, 1955; see also Hintikka, 1985).

7 There are logical systems which eliminate quantifiers, for example, *combinatory logic* (see Curry's and Feys' (1958) and Fine's (1985) theory of arbitrary objects. Perhaps a translation between these systems and mental models may provide the desired results.

8 This is simply due to the inability to draw more than four overlapping two-dimensional shapes such that all possible relationships between them are represented.

9 This is far less important than *soundness*. However, if mental-models theory is to avoid the charge of *ad hoc* extension to deal with new phenomena, then some account of *expressiveness* must be provided. Otherwise there can be little confidence that the notation is sufficiently well understood to perform the functions demanded of it.

10 At the beginning of Johnson-Laird and Byrne (1991) the example of a classic piece of default reasoning by Sherlock Holmes is provided which eloquently illustrates this point.

11 The cost is that logical negation is not fully implemented in such a database.

12 These possibilities are known as different *extensions* of a default theory. A default theory is simply a collection of axioms, including at least one default rule, which describes the behaviour of a particular domain.

13 It also suggests that sensible reasoning in novel domains does not demand an abstract inferential competence sensitive to the logical form of arguments. Just as with precedent, old judgements are brought to bear on new problems.

REFERENCES

Baars, B.J. (1986) *The Cognitive Revolution in Psychology*, New York: Guilford Press.

Baddeley, A.D. (1986) *Working Memory*, Oxford: Clarendon Press.

Beth, E.W. (1955) 'Semantic entailment and formal derivability', *Mededelingen van de Koninklijke Nederlande Akadamie van Wetenschappen, Afd. Letterkunde* 18: 309–42.

Braine, M.D.S. (1978) 'On the relationship between the natural logic of reasoning and standard logic', *Psychological Review* 85: 1–21.

Braine, M.D.S., Reiser, B.J., and Rumain, B. (1984) 'Some empirical justification for a theory of natural propositional logic', *The Psychology of Learning and Motivation*, vol. 18, New York: Academic Press.

Bransford, J.D. and Johnson, M. (1972) 'Contextual prerequisites for understanding: some investigations of comprehension and recall', *Journal of Verbal Learning and Verbal Behaviour* 11: 717–26.

—— (1973) 'Considerations of some problems of comprehension', in W.G. Chase (ed.) *Visual Information Processing*, New York: Academic Press, pp. 389–92.

Bransford, J.D. and McCarrell, N.S. (1975) 'A sketch of a cognitive approach to comprehension: some thoughts on what it means to comprehend', in W.B. Weimer and D.S. Palermo (eds) *Cognition and Symbolic Processes*, Hillsdale, NJ: Erlbaum, pp. 189–229.

Bransford, J.D., Barclay, J.R., and Franks, J.J. (1972) 'Sentence memory: a constructive versus interpretive approach', *Cognitive Psychology* 3: 193–209.

Brown, H.I. (1988) *Rationality*, London: Routledge.

Byrne, R.M.J. (1989) 'Suppressing valid inferences with conditionals', *Cognition* 31: 1–21.

Charniak, E. and McDermott, D. (1985) *An Introduction to Artificial Intelligence*, Reading, MA: Addison-Wesley.

Chater, N. (1989) *Learning to Respond to Structure in Time*, Research Initiative in Pattern Recognition Technical Report, Malvern: RSRE September.

Chater, N. and Oaksford, M. (1990) 'Autonomy, implementation and cognitive architecture: a reply to Fodor and Pylyshyn', *Cognition* 34: 93–107.

Cheng, P.W. and Holyoak, K.J. (1985) 'Pragmatic reasoning schemas', *Cognitive Psychology* 17: 391–416.

—— (1989) 'On the natural selection of reasoning theories', *Cognition* 33: 285–313.

Cherniak, C. (1986) *Minimal Rationality*, Cambridge, MA: MIT Press.

Churchland, P.M. and Churchland, P.S. (1983) 'Stalking the wild epistemic engine', *Nous* 17: 5–18.

Clark, H.H. (1977) 'Bridging' in P.N. Johnson-Laird and P.C. Wason (eds) *Thinking: Readings in Cognitive Science*, Cambridge: Cambridge University Press, pp. 411–20.

Cook, S. (1971) 'The complexity of theorem proving procedures', in *The Third Annual Symposium on the Theory of Computing*, New York, pp. 151–8.

Cosmides, L. (1989) 'The logic of social exchange: has natural selection shaped how humans reason? Studies with the Wason selection task', *Cognition* 31: 187–276.

Cummins, D.D., Lubart, T., Alksnis, O., and Rist, R. (1991) 'Conditional reasoning and causation', *Memory & Cognition* 19: 274–82.

Curry, H.B. (1956) *An Introduction to Mathematical Logic*, Amsterdam: Van Nostrand.

Curry, H.B. and Feys, R. (eds) (1958) *Combinatory Logic*, Amsterdam: North-Holland.

Elman, J.L. (1988) *Finding Structure in Time*, CRL Technical Report 8801, San Diego: Centre for Research in Language, University of California.

Evans, J. St B.T. (ed.) (1983a) 'Selective processes in reasoning', *Thinking and Reasoning: Psychological Approaches*, London: Routledge & Kegan Paul.

—— (1983b) 'Linguistic determinants of bias in conditional reasoning', *Quarterly Journal of Experimental Psychology* 35A: 635–44.

—— (1984) 'Heuristic and analytic processes in reasoning', *British Journal of Psychology* 75: 451–68.
—— (1989) *Bias in Human Reasoning: Causes and Consequences,* London: Erlbaum.
—— (1991) 'Theories of human reasoning: the fragmented state of the art', *Theory & Psychology* 1: 83–105.
Fine, K. (1985) *Reasoning with Arbitrary Objects*, Oxford: Basil Blackwell.
Fodor, J.A. (1983) *Modularity of Mind*, Cambridge MA: MIT Press.
—— (1987) *Psychosemantics: The Problem of Meaning in the Philosophy of Mind*, Cambridge, MA: MIT Press.
Fodor, J.A. and Pylyshyn, Z.W. (1981) 'How direct is visual perception? Some reflections on Gibson's "Ecological Approach"', *Cognition* 9: 139–96.
—— (1988) 'Connectionism and cognitive architecture: a critical analysis', *Cognition* 28: 3–71.
Garey, M.R. and Johnson, D.S. (1979) *Computers and Intractability: A Guide to the Theory of NP-Completeness*, San Francisco: W.H. Freeman.
Gentzen, G. (1934) 'Untersuchungen über das logishce Schliessen', *Mathematische Zeitschrifft* 39: 176–210.
Ginsberg, M.L. (ed.) (1987) *Readings in Nonmonotonic Reasoning*, Los Altos, CA: Morgan Kaufman.
Glymour, C. (1987) 'Android epistemology and the frame problem: comments on Dennett's "Cognitive Wheels"', in Z.W. Pylyshyn (ed.) *The Robot's Dilemma: The Frame Problem in Artificial Intelligence*, Norwood, NJ: Ablex, pp. 65–76.
Glymour, C. and Thomason, R.H. (1984) 'Default reasoning and the logic of theory perturbation', unpublished manuscript, History and Philosophy of Science Department, University of Pittsburgh.
Gregory, R.L. (1977) *Eye and Brain*, 3rd edn, London: Weidenfeld & Nicolson.
Harman, G. (1986) *Change in View*, Cambridge, MA: MIT Press.
Henle, M. (1962) 'On the relation between logic and thinking', *Psychological Review* 69: 366–78.
Hintikka, J. (1955) 'Form and content in quantification theory', *Acta Philosophica Fennica* 8: 11–55.
—— (1985) 'Mental models, semantical games, and varieties of intelligence', unpublished manuscript, University of Florida.
Hodges, W. (1975) *Logic*, Harmondsworth: Penguin.
Hogger, C.J. (1984) *An Introduction to Logic Programming*, London: Academic Press.
Horowitz, E. and Sahni, S. (1978) *Fundamentals of Computer Algorithms*, Rockville, Maryland: Computer Science Press.
Inhelder, B. and Piaget, J. (1958) *The Growth of Logical Reasoning*, New York: Basic Books.
James, W. (1950) *The Principles of Psychology*, vol. 1, New York: Dover (originally published in 1890).
Johnson-Laird, P.N. (1975) 'Models of deduction', in R.J. Falmagne (ed.) *Reasoning: Representation and Process*, Hillsdale, NJ: Erlbaum.
—— (1983) *Mental Models: Towards a Cognitive Science of Language, Inference and Consciousness*, Cambridge: Cambridge University Press.
—— (1986) 'Reasoning without logic', in T. Myers, K. Brown, and B. McGonigle (eds) *Reasoning and Discourse Processes*, London: Academic Press, pp. 13–50.
Johnson-Laird, P.N. and Byrne, R.M.J. (1991) *Deduction*, Hillsdale, NJ: Erlbaum.
Jordan, M.I. (1986) *Serial Order: A Parallel Distributed Approach*, Institute for Cognitive Science Report 8604, San Diego: University of California.

Kahneman, D., Slovic, P., and Tversky, A. (eds) (1982) *Judgement Under Uncertainty: Heuristics and Biases*, Cambridge: Cambridge University Press.
Kuhn, T.S. (1962) *The Structure of Scientific Revolutions*, Chicago: University of Chicago Press.
Lakatos, I. (1970) 'Falsification and the methodology of scientific research programmes', in I. Lakatos and A. Musgrave (eds) *Criticism and the Growth of Knowledge*, Cambridge: Cambridge University Press, pp. 91–196.
Levesque, H.J. (1985) 'A fundamental tradeoff in knowledge representation and reasoning', in R.J. Brachman and H.J.Levesque (eds) *Readings in Knowledge Representation*, Los Altos, CA: Morgan Kaufman.
—— (1988) 'Logic and the complexity of reasoning', *Journal of Philosophical Logic* 17: 355–89.
Lewis, D. (1973) *Counterfactuals*, Oxford: Oxford University Press.
Loui, R.P. (1987) 'Response to Hanks and McDermott: temporal evolution of beliefs and beliefs about temporal evolution', *Cognitive Science* 11: 283–97.
McArthur, D.J. (1982) 'Computer vision and perceptual psychology', *Psychological Bulletin* 92: 283–309.
McCarthy, J.M. and Hayes, P. (1969) 'Some philosophical problems from the standpoint of artificial intelligence', in B. Meltzer and D. Michie (eds) *Machine Intelligence* 4, New York: Elsevier.
McDermott, D. (1986) *A Critique of Pure Reason*, Technical Report, Department of Computer Science, Yale University, June, 1986.
Markovits, H. (1984) 'Awareness of the "possible" as a mediator of formal thinking in conditional reasoning problems', *British Journal of Psychology* 75: 367–76.
—— (1985) 'Incorrect conditional reasoning among adults: competence or performance', *British Journal of Psychology* 76: 241–7.
Medin, D.L. and Schaffer, M.M. (1978) 'Context theory of classification learning', *Psychological Review* 85: 201–38.
Miller, G.A. (1956) 'The magical number 7±2: some limits on our capacity for processing information', *Psychological Review* 63: 81–97.
Minsky, M. (1975) 'Frame-system theory', in R. Schank and B.L. Nash-Webber (eds) *Theoretical Issues in Natural Language Processing*, Cambridge, MA, 10–13 June 1975.
Mynatt, C.R., Doherty, M.E., and Tweney, R.D. (1977) 'Confirmation bias in a simulated research environment: an experimental study of scientific inference', *Quarterly Journal of Experimental Psychology* 29: 85–95.
Nosofsky, R.M. (1986) 'Attention, similarity and the identification-categorisation relationship', *Journal of Experimental Psychology: General* 115: 39–57.
Oaksford, M. and Chater, N. (1991) 'Against logicist cognitive science', *Mind & Language* 6: 1–38.
—— (1992) 'Bounded rationality in taking risks and drawing inferences', *Theory & Psychology* 2: 225–30.
—— (forthcoming) *Cognition and Inquiry*, London: Academic Press.
Oaksford, M. and Stenning, K. (1992) 'Reasoning with conditionals containing negated constituents', *Journal of Experimental Psychology: Learning, Memory & Cognition* 18: 834–54.
Oaksford, M., Chater, N., and Stenning, K. (1990) 'Connectionism, classical cognitive science and experimental psychology', *AI & Society* 4: 73–90. Also in A. Clark and R. Lutz (eds) (1992) *Connectionism in Context*, Berlin: Springer-Verlag, pp. 57–74.
Osherson, D. (1975) 'Logic and models of logical thinking', in R.J. Falmagne (ed.) *Reasoning: Representation and Process*, Hillsdale, NJ: Erlbaum.

Pylyshyn, Z.W. (ed.) (1987) *The Robot's Dilemma: The Frame Problem in Artificial Intelligence*, Norwood, NJ: Ablex.
Quine, W.O. (1959) *Methods of Logic*, New York: Holt, Rinehart, & Winston.
Reiter, R. (1980) 'A logic for default reasoning', *Artificial Intelligence* 13: 81–132.
—— (1985) 'On reasoning by default', in R. Brachman and H. Levesque (eds) *Readings in Knowledge Representation*, Los Altos, CA: Morgan Kaufman (originally published in 1978).
Rips, L.J. (1983) 'Cognitive processes in propositional reasoning', *Psychological Review*, 90: 38–71.
Rohwer, R. (1990) 'The "Moving Targets" training algorithm', in L.B.Almeida and C.J. Wellekens (eds) *Lecture Notes in Computer Science 412: Neural Networks*, Berlin: Springer-Verlag, pp. 100–9.
Rosch, E. (1973) 'On the internal structure of perceptual and semantic categories', in T. Moore (ed.) *Cognitive Development and the Acquisition of Language*, New York: Academic Press.
—— (1975) 'Cognitive representation of semantic categories', *Journal of Experimental Psychology: General* 104: 192–233.
Rumain, B., Connell, J., and Braine, M.D.S. (1983) 'Conversational comprehension processes are responsible for reasoning fallacies in children as well as adults. IF is not the biconditional', *Developmental Psychology* 19: 471–81.
Rumelhart, D.E. (1989) 'Toward a microstructural account of human reasoning', in S. Vosnaidou and A. Ortony (eds) *Similarity and Analogical Reasoning*, Cambridge: Cambridge University Press, Ch. 10, pp. 298–312.
Rumelhart, D.E., Smolensky, P., McClelland, J.L., and Hinton, G.E. (1986) 'Schemata and sequential thought processes in PDP models', in J.L. McClelland and D.E. Rumelhart (eds) *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, vol 2: Psychological and Biological processes*, Cambridge, MA: MIT Press, Ch. 14, pp. 7–57.
Simon, H.A. (1969) *The Sciences of the Artificial*, Cambridge, MA: MIT Press.
Stalnaker, R. (1968) 'A theory of conditionals', in N. Rescher (ed.) *Studies in Logical Theory*, Oxford: Oxford University Press.
Stenning, K. and Oaksford, M. (1989) *Choosing Computational Architectures for Text Processing*, Technical Report No. EUCCS/RP-28, Edinburgh: Centre for Cognitive Science, University of Edinburgh, April, 1989.
Taplin, J.E. (1971) 'Reasoning with conditional sentences', *Journal of Verbal Learning and Verbal Behaviour* 10: 219–25.
Taplin, J.E. and Staudenmayer, H. (1973) 'Interpretation of abstract conditional sentences in deductive reasoning', *Journal of Verbal Learning and Verbal Behaviour* 12: 530–42.
Tsotsos, J.K. (1990) 'Analyzing vision at the complexity level', *Behavioral & Brain Sciences* 13: 423–69.
Tversky, A. and Kahneman, D. (1974) 'Judgement under uncertainty: heuristics and biases', *Science* 185: 1124–31.
Tukey, D.D. (1986) 'A philosophical and empirical analysis of subjects' modes of inquiry in Wason's 2–4–6 task', *Quarterly Journal of Experimental Psychology* 38A: 5–33.
Tweney, R.D. (1985) 'Faraday's discovery of induction: a cognitive approach', in D. Gooding and F. James (eds) *Faraday Rediscovered*, London: Macmillan, pp. 159–209.
Wason, P.C. (1965) 'The contexts of plausible denial', *Journal of Verbal Learning and Verbal Behavior* 4: 7–11.
—— (1966) 'Reasoning', in B. Foss (ed.) *New Horizons in Psychology*, Harmondsworth: Penguin.