THINKING AND REASONING, 1995, *1*, (2), 121–152

Theories of Reasoning and the Computational Explanation of Everyday Inference

Mike Oaksford
University of Warwick, UK

Nick Chater
University of Oxford, UK

Following Marr (1982), any computational account of cognition must satisfy constraints at three explanatory levels: computational, algorithmic, and implementational. This paper focuses on the first two levels and argues that current theories of reasoning cannot provide explanations of everyday defeasible reasoning, at either level. At the algorithmic level, current theories are not computationally tractable: they do not "scale-up" to everyday defeasible inference. In addition, at the computational level, they cannot specify *why* people behave as they do both on laboratory reasoning tasks and in everyday life (Anderson, 1990). In current theories, logic provides the computational-level theory, where such a theory is evident at all. But logic is not a descriptively adequate computational-level theory for many reasoning tasks. It is argued that better computational-level theories can be developed using a probabilistic framework. This approach is illustrated using Oaksford and Chater's (1994) probabilistic account of Wason's selection task.

INTRODUCTION

In this paper, we argue that current theories of reasoning are unable to provide computational explanations of everyday human reasoning. As in other areas of cognitive psychology, computational ideas feature prominently in reasoning research. Current reasoning theories, for example, use ideas from heuristic search (Evans, 1984, 1989; Newell & Simon, 1972), theorem proving (Newell & Simon, 1972; Rips, 1983, 1994), and frame system theory (Cheng & Holyoak, 1985; Minsky, 1975; Rumelhart, 1980) to explain reasoning performance. Further, Johnson-Laird (1983) has argued that cognitive theories should be

Requests for reprints should be sent to Mike Oaksford, Department of Psychology, University of Warwick, Coventry CV4 7AL, UK. Email: pysad@csv.warwick.ac.uk or chater@psy.ox.ac.uk

We thank David O'Brien, Phil Johnson-Laird, and Ken Manktelow for their helpful comments on an earlier version of this article.

sufficiently precise to implement in a computer program. In short, reasoning theorists have adopted the view that reasoning can be explained in computational terms.

What is it to give a computational explanation? The most influential account of computational explanation in cognitive science is due to Marr (1982; and see Anderson, 1990: pp.4–5, for a summary of other accounts). Marr defined three levels of computational explanation. At the *computational* level (Marr, 1982, p.24) "the performance of the device is characterised as a mapping from one kind of information to another, the abstract properties of this mapping are defined precisely, and its appropriateness and adequacy for the task at hand are demonstrated". Marr uses the example of a cash register. The theory of arithmetic provides the computational-level analysis of this device. Demonstrating its appropriateness involves showing that our intuitive constraints on the operation of a cash register map directly onto this mathematical theory (Marr, 1982, p.22). Anderson (1990) refers to the computational level as the "rational" level—providing a computational-level analysis of some task performance amounts to specifying the rational function of the observed behaviour on that task. Marr (1982, p.27) viewed "the computational level . . . [as] critically important from an information processing point of view". For Marr, trying to understand a computational process without this level of analysis is like trying to understand bird flight without a theory of aerodynamics. Thus Marr took the computational level to be the logically prior starting point for providing computational explanations.

The *algorithmic* level describes how to compute the function specified at the computational level. This level also involves specifying the representations that the algorithm manipulates in computing the function. Thus in the case of the cash register, using Arabic numerals as the representational notation involves (Marr, 1982, p.22) using the standard rules "about adding the least significant digits first and 'carrying' the sum if it exceeds 9" as an algorithm. Although the choice of representation constrains the choice of algorithm, it is not uniquely constrained—there may be several ways of computing a certain function using the same representation. An important constraint on the choice of algorithm is its computational efficiency (Marr, 1982, p.24): "which [algorithm] is chosen will usually depend on any particularly desirable or undesirable characteristics that the algorithm may have; for example, one algorithm may be much more efficient than another." We argue later on that a major problem for current reasoning theories is that they attempt to apply demonstrably inefficient algorithms to modes of reasoning that people appear to perform very efficiently.

The *implementational* level outlines the physical realisation of the algorithm. This level involves the detailed physical structure—the computer architecture—that implements the algorithm. We will have little to say about this level in this paper. It is worth noting, however, that this level also constrains the choice of algorithm. For example, as we have observed elsewhere (Chater & Oaksford,

1990), classical symbolic algorithms are unlikely to run efficiently on connectionist hardware. We shall use Marr's levels of computational explanation to re-evaluate current theories in the psychology of reasoning.

We evaluate these theories for their ability to account for everyday human inference. We observe later that a crucial difference between everyday inference and deductive reasoning is that everyday inference is defeasible: conclusions only follow tentatively, rather than certainly, from premises. Furthermore, we suggest that everyday inference involves large numbers of premises embodying world knowledge, rather than the very small number of premises generally used in laboratory reasoning experiments. We argue that these features of everyday inference raise difficult problems for current reasoning theories at both the algorithmic and computational levels. We suggest that problems at the prior computational level can be alleviated by using probability theory rather than logic to model uncertain reasoning. Problems at the algorithmic level remain a difficult, but little acknowledged, challenge to all theories of reasoning.

We have organised this paper as follows. We first introduce the four main theories of reasoning, and discuss whether they generalise to everyday reasoning. To generalise successfully, these theories must be adequate at both the algorithmic and computational levels. We then consider each of these levels in turn. In *The Algorithmic Level,* we outline computational complexity theory and show how it applies to theories of everyday reasoning in artificial intelligence. We then show that no current theory of reasoning can provide an efficient or "tractable" algorithm for defeasible reasoning. In *The Computational Level,* we discuss why a computational-level theory must be both normatively justified and descriptively adequate. We then show that logic-based accounts of everyday, defeasible inference are descriptively inadequate. We further argue that because logic provides the only computational theory used in reasoning research, current reasoning theories are unable to generalise to everyday inference. We deal with several attempts to defend logic-based approaches, which deny that everyday inference is invariably defeasible, arguing that none of these attempts is successful. Finally, we suggest that because everyday reasoning is uncertain, we should seek appropriate computational-level theories using the mathematical calculus of uncertainty—probability theory. We illustrate this approach using our recent probabilistic computational-level theory of Wason's selection task (Oaksford & Chater, 1994).

THEORIES OF REASONING AND THEIR GENERALITY

Evans (1991) offers a four-way classification of deductive reasoning theories and a three-way characterisation of the questions they must try to answer. The questions that a reasoning theory must address are: the competence question—how do subjects often solve reasoning problems?; the bias question—why do subjects also make many systematic errors?; and the content and context

question—why is it that the content and context of a problem influence inferential performance?. Evans (1991) argues that the four theories of reasoning concentrate on one question or the other, but none provides an account of all three.

Evans notes that two theories concentrate on the competence question. The *mental logic* approach argues that people reason using formal inference rules such as *modus ponens* (given *if p, then q* and *p* you can infer *q*) that rely only on the syntactic form of the premises (Braine, 1978; Henle, 1962; Inhelder & Piaget, 1958; Johnson-Laird, 1975; Osherson, 1975; Rips, 1983). *Mental models* theory argues that people base their reasoning on semantic principles (Johnson-Laird, 1983; Johnson-Laird & Byrne, 1991). Johnson-Laird and Byrne (1991), for example, argue that the complexity of inference in mental models matches the inferential difficulty subjects experience in laboratory tasks. We argue that these two theories are the only current reasoning theories that specify a computational-level theory—this theory is formal logic.

Evans notes that two further theories concentrate on content effects and errors and biases in reasoning. *Pragmatic reasoning schema* theory proposes inference rules 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. These tasks use contentful rules about permission relations, e.g. *if you are drinking alcohol, you must be over 18 years of age*. Lastly the *heuristic approach* explains systematic errors and biases by people using various short-cut processing strategies (Evans, 1983, 1984, 1989). We argue that these two theories do not use formal logic as their computational-level theory. Pragmatic reasoning schemas are grounded in intuitively plausible rules rather than in a formal, computational-level theory.[1] The heuristic approach by definition can only supplement a computational-level theory.

All four theories of deductive reasoning account for performance on a relatively narrow range of laboratory reasoning tasks. Do these accounts *generalise* to inferential processes in everyday reasoning? If not, then the psychology of deductive reasoning would be of no more general interest than the psychology of crossword puzzles. But clearly psychologists of reasoning intend their theories to generalise to everyday inference. Rips (1994) is explicit on this, in his exploration of what he calls the *deduction system hypothesis*—that logic is central to cognition because it underlies many other cognitive abilities. Johnson-Laird and Byrne (1991, pp.2–3) focus on deduction:

> . . . because of its intrinsic importance: it plays a crucial role in many tasks. You need to make deductions in order to formulate plans and to evaluate actions; to

[1]Within pragmatic reasoning schema theory, there are proposals to use ideas from jurisprudence to provide computational-levels accounts of certain kinds of reasoning contracts (Holyoak & Cheng, in press).

> determine the consequences of assumptions and hypotheses; to interpret and formulate instructions, rules and general principles; to pursue arguments and negotiations; to weigh evidence and to assess data; to decide between competing theories; and to solve problems.

How could we find out whether deductive reasoning really does underlie performance across this wide range of everyday tasks? It is extremely difficult to imagine how you might empirically test such a claim. However, because we are working on the assumption that reasoning is a computational process, it should be possible to construct computational models of such processes, based on logical inference. Fortunately, this is not merely an interesting possible line of future inquiry—the attempt to model diverse areas of cognition using logical methods has been the central goal of artificial intelligence (AI) since its inception. We shall, therefore, draw conclusions from attempts in AI to model everyday tasks using logical methods.

In AI, cases where the human cognitive system far outstrips the capacities of computational memory systems (Oaksford, Chater, & Stenning, 1990) have been of particular interest. AI programs suffer from a well-documented limitation on retrieval from long-term memory. McCarthy and Hayes (1969) called this limitative finding the "frame problem" (see Pylyshyn, 1987 for overviews). Glymour (1987, p.65) characterises the frame problem as follows: "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?". The frame problem may arise for any task requiring the deployment of prior world knowledge.

In order to generalise to everyday inference, theories of deductive reasoning must confront the frame problem. We now argue that everyday inference is defeasible or non-monotonic: that is, the addition of further premises can defeat previous conclusions. The problems of non-monotonic reasoning give rise to the frame problem. Current theories of reasoning typically do not directly attempt to capture non-monotonic reasoning, and hence do not appear even to attempt to generalise to everyday inference.

The problems of non-monotonic reasoning arise at both the computational and algorithmic levels of explanation. At the computational level, we require an account of *what* inferences people draw; and at the algorithmic level, we require an account of *how* they draw those inferences. Current reasoning theories have concentrated on developing algorithmic-level accounts, and either provide no formal computational-level theory or use logic in this role. In the next section, *The Algorithmic Level*, we argue that these algorithmic theories cannot generalise to everyday reasoning because they are computationally intractable, and that these intractability problems are especially acute if generalised to everyday reasoning. In the following section, *The Computational Level*, we suggest that some of these problems may derive from using inappropriate computational-level theories of everyday inference. We argue that probability

theory may provide more appropriate computational-level theories, although it does not resolve problems at the algorithmic level.

THE ALGORITHMIC LEVEL

This section has three parts. First, we outline computational complexity theory, which characterises the time and space requirements of algorithms for solving particular problems, independent (within very wide limits) of the computational device used. Second, we consider how this theory applies to everyday inference, and show that approaches to everyday inference are computationally infeasible. Third, we argue that these problems apply to each of the four contemporary approaches to reasoning, when generalised to everyday inference.

Computational Complexity Theory

Psychologists have always been concerned with real-time processing. Indeed, in many areas of psychology reaction times have been the primary source of constraint (e.g. Posner, 1978). Modelling precise real-time characteristics of inference has been of less concern to psychologists of reasoning. Nonetheless, it is uncontroversial that any psychologically plausible algorithmic account of human reasoning must be consistent with human inference happening in real time. However, when generalised to handle common-sense inference, the predictions of reasoning theories regarding the time course of inference may be unacceptable. Specifically, these theories may have to predict that people cannot complete the simplest everyday inference within a human life-time, let alone within the time available for real human inference.

How can we make predictions about real-time performance of an algorithm running on the computational machinery of the human brain? *Prima facie*, this appears to require a detailed knowledge of how to implement cognitive algorithms in the brain, information that is simply not available. However, computer science has shown that the broad pattern of real-time performance, although not its detailed time-course, is predictable without knowledge of the underlying computational hardware, or of how to implement the algorithm in that hardware. *Computational complexity theory* (see for example, Garey & Johnson, 1979; Horowitz & Sahni, 1978) classifies algorithms depending on how their time and space requirements increase with the length of the input. For human reasoning the length of the input corresponds to the number of premises in an argument or facts in a knowledge-base.

We only need a few of the results from computational complexity theory to allow us to show in the next section how theories of everyday reasoning must confront the frame problem. Computational complexity theory divides problems into two main classes—those that have a *tractable* algorithm for their solution and those that do not. An intractable algorithm is one where the time and space requirements grow as an exponential function of the length of the input (number

of premises). Computer scientists generally regard algorithms for which this function is a polynomial (or less) as tractable. There is some uncertainty as to which algorithms fall into the intractable class. For one class of problems (called *NP-complete*, see Appendix 1) all known algorithms are exponential and hence intractable. Cook (1971) has shown how to characterise all these problems in the same way—they are all problems of logical consistency checking. Cook's theorem states that if one of these problems has a polynomial time algorithm then they all have. However, because—as a matter of fact—no one has managed to devise a polynomial time algorithm for any of these problems, computer scientists have assumed that they are generally intractable (this is usually stated as the conjecture that NP $\neq$ P). (See Appendix 1 for relevant technical details).

Everyday Inference

In this section, we observe how computational complexity poses problems for theories of everyday inference. Although the psychology of reasoning has rarely considered computational complexity, it has been a central concern in other areas of cognitive science. For example, early work on bottom-up object recognition of block 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). Furthermore, researchers into risky decision making realised very early on that complexity issues were relevant. Probabilistic inference makes exponentially increasing demands on computational resources even for problems involving very moderate amounts of information (Charniak & McDermott, 1985).

Although the psychology of reasoning focuses on deduction, according to which conclusions follow with certainty if the premises are true, most human inference is uncertain. For example, activities such as text comprehension, classification, categorisation, and perception all rely on inferential processes, which are *defeasible*—subsequent information may *defeat* earlier conclusions (Oaksford & Chater, 1991, 1992, 1993). For example, on learning that *Tweety is a bird* you may elaboratively infer that *Tweety can fly*. However, the common sense generalisation—*all birds can fly*—that licenses this inference is defeasible. When you subsequently learn that *Tweety is an ostrich* this defeats the conclusion that *Tweety can fly*.

Using this example, we can now illustrate how complexity issues arise for logic-based AI approaches to everyday, defeasible reasoning (McDermott, 1987). The standard logical approach (e.g. Reiter, 1980, 1985) is to propose a "closed world" assumption—an AI program bases its inferences on the "closed world" consisting of the current contents of its data-base, where the contents of the data-base provide additional premises. Reiter would treat our example rule, "all birds can fly" as meaning that *If x is a bird, and there is no reason to suppose otherwise, then x can fly*. So when you learn that *Tweety is a bird*, and

you cannot generate a counter-example from your current data-base, i.e. you can not generate a reason to suppose that *Tweety cannot fly*, then it is reasonable for you to infer that *Tweety can fly*. This means that every time a reasoner draws a conclusion from a default rule they must exhaustively search the whole of their data-base to ensure that no counter-example is available. This is equivalent to checking the consistency of the data-base, which, as we noted earlier means that it is computationally intractable.

These considerations imply that logic-based approaches to defeasible reasoning can only apply to small data-bases, i.e. to small sets of premises, before complexity bites. For psychologists of reasoning, this may not appear to be a problem (Garnham, 1993; see Chater & Oaksford, 1993). In most explicit reasoning tasks, the premise sets are extremely small, usually consisting of two or three premises. And in such inference tasks, it is true that increases in the explicit premises beyond this number produces catastrophic performance breakdown (Johnson-Laird, 1983). At this point, the contrast between laboratory-based explicit reasoning tasks, and real-time everyday inference becomes critical. For in everyday inference, the subject does not reason over a handful of premises specified by the experimenter, but rather must make the best inference possible given their pre-existing data-base consisting of large amounts of world knowledge.

From the point of view of complexity constraints, the crucial question is how many premises are involved in typical everyday defeasible inference. Considerations from AI and psychology are relevant here. In AI, the formalisation of the tiniest fragment of world knowledge in logical terms involves enormous numbers of premises (e.g. Hayes, 1978, 1984a, 1984b). Furthermore, the interconnected character of world knowledge indicates that knowledge about some specific domain cannot be perfectly isolated from knowledge from other domains (Fodor, 1983). It is for this reason that schema theorists no longer assume that they can isolate schemas from one another, as independent data-bases, but rather they propose that schemata must be richly interconnected (Rumelhart, 1980; Schank, 1982). Thus, attempts to formalise knowledge within AI suggest that the number of premises in the data-base relevant to a defeasible inference is very large, and indeed, probably includes the whole of world knowledge. Furthermore, the psychological evidence points in the same direction. For example, the rapidly drawn elaborative inferences (which are uncontroversially defeasible) involved in understanding the simplest of texts draw on large amounts of world knowledge (Clark, 1977; Garrod & Sanford, 1977; Kintsch & van Dijk, 1978; O'Brien, Shank, Myers, & Rayner, 1988). So, although computational complexity constraints may not bite for explicit laboratory reasoning tasks, they do for accounts of everyday reasoning.

Thus, if psychologists of reasoning intend their accounts of laboratory reasoning tasks to generalise to everyday reasoning, they face a paradox: although they can account for people's poor performance on explicit reasoning

tasks, it appears that they cannot account for how everyday reasoning is possible at all. In the next section we confirm this concern, considering each of the four classes of reasoning theory in turn.

Theories of Reasoning and Computational Complexity

We deal with the four theories of reasoning in the order in which we introduced them: mental logic, mental models, pragmatic reasoning schemas, and the heuristic approach.

Mental Logic. The contemporary mental logic view explains explicit reasoning performance by appeal to various natural deduction systems (Gentzen 1934) with (Rips 1983) or without (Braine 1978) an account of the control processes that animate the inference rules. From a complexity-theoretic standpoint, mental logic seems unpromising. Even for standard monotonic logic, the general problem of deciding whether a given finite set of premises logically implies a particular conclusion is computationally intractable (Cook, 1971).[2] Moreover, the complexity results we discussed earlier derived from logical attempts to account for default reasoning in AI knowledge representation. Consequently a mental logic is unlikely to generalise to everyday defeasible reasoning.

Mental Models. Logic's failure to generalise to everyday inference appears to add further weight to the mental modeller's claim that "there is no mental logic". On the mental models view, semantic methods of proof should replace the syntactic formalisms of the mental logician (e.g. Johnson-Laird 1983; Johnson-Laird & Byrne, 1991). However, recently, mental modellers have tempered their claim that "there is no mental logic". For example (Johnson-Laird & Byrne, 1991, p.212): ". . . 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 counter-examples)". So the dispute is not about *whether* there is a mental logic, but about *how* to implement it in the mind. Note also that the problem of searching for counter-examples, which is the engine of the mental models approach, is no more or less than the problem of consistency checking. Specifically, the problem of finding a counter-example is the problem of finding a case where the premises are true and the conclusion is false; this will be possible if and only if the negation of the conclusion is consistent with the premises (Enderton, 1972). Thus, searching for counter-examples just is consistency checking. This identity appears to refute immediately the possibility that mental models offers a tractable approach to everyday inference.

---

[2]This applies equally to semantic proof procedures, such as truth tables and semantic tableaux, as to syntactic procedures such as axioms or natural deduction systems.

Mental models theorists are not unaware of this problem and argue that using *arbitrary exemplars* may allow mental models theory to develop a tractable proof procedure (Johnson-Laird, 1983). However, there are no complexity results for the algorithms that manipulate mental models. In the absence of such results there is no evidence that mental models could fare any better than mental logic in providing computationally tractable algorithms for everyday inference.

Pragmatic Reasoning Schema Theory. Pragmatic reasoning schema theory emphasises the role of domain-specific knowledge in reasoning tasks (Cheng & 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. They invoke permission schemas to explain the results from some thematic versions of Wason's selection task where the rule determines whether an agent may perform a particular action. 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 & Holyoak, 1989, for a critique).

If the domains over which the search for counter-examples were suitably constrained, then exhaustive searches may be feasible. However, as we noted earlier, AI researchers have made extensive use of schema theories and have found that they run directly into the frame problem (Fodor, 1983; Pylyshyn, 1987). Indeed Hayes (1979) has shown that early schema theories are equivalent to logical formalisms; and Reiter (1985) has re-characterised the way schema theories handle defaults using non-monotonic logic. Although schema theories may prove useful in describing performance in laboratory experiments, AI researchers have tried, tested, and abandoned them as computationally tractable accounts of everyday defeasible inference.

Heuristic Approaches. Only the heuristic approach (Evans, 1983, 1984, 1989) explicitly addresses the issue of cognitive limitations. In computer science the use of heuristics may render a computationally intractable problem manageable. Using a generally intractable algorithm with a heuristic can provide tractable, approximate solutions for many problem instances (Horowitz & Sahni, 1978). You trade accuracy for speed. In this section we observe that the heuristic approach does not address the issue of intractability.

Evans (1991) notes that the heuristic approach is *not* an approach to human reasoning in its own right—it can only supplement a theory of competence such as mental logic or mental models. Thus the viability of the heuristic approach depends in this context on whether there are heuristics that can allow reasonably reliable consistency checking over data-bases of a cognitively realistic size. Although this remains a possibility, it has so far eluded researchers in AI and

computer science. Hence there are currently no grounds to believe that a heuristic approach is viable in general, and certainly there are no specific heuristics proposed within the psychology of reasoning that could resolve the problem of computational intractability.

More recently Evans (in press) has proposed that *relevance* is crucial to human reasoning. He suggests that the heuristics he has proposed serve to retrieve relevant information from memory. As we noted earlier (see quote from Glymour, 1987), retrieving relevant information from memory is just the frame problem. So relevance approaches (see also, Sperber, Cara, & Girotto, in press), although implicitly conceding that reasoning theories must confront the frame problem (which is a step forward), can do nothing to resolve this problem (Oaksford & Chater, 1991).

These algorithmic level problems are most pressing for theories of reasoning based on logic; but they are equally serious for the probabilistic approach that we shall advocate later. Hence, although we hope to have established that complexity considerations should be of serious concern for psychological theories of reasoning, we do not take them to militate decisively against current theories of reasoning. The more fundamental difficulties for current theories of reasoning come at the computational level; and it is here that the probabilistic approach is most promising.

THE COMPUTATIONAL LEVEL

In this section we first discuss what is required of a computational-level theory. We then argue that only two theories—mental logic and mental models—embody a computational-level theory, and that in both cases logic provides this theory. We then show that logic provides a completely inappropriate framework for modelling everyday defeasible inference, which suggests that mental logics and mental models can not generalise to deal with everyday inference. We then consider possible responses to this line of argument from advocates of each approach, and argue that these responses are inadequate. We will then argue that an appropriate computational-level theory that captures the uncertain character of everyday inference should be provided not by logic but by the calculus of uncertain reasoning—probability theory. We illustrate this approach using our probabilistic computational-level theory of Wason's selection task (Oaksford & Chater, 1994).

Computational level theories

As we outlined in the introduction, according to Marr (1982) a computational-level theory specifies *what* is computed and *why* in the performance of some task. Marr uses the example of a cash register where the theory of addition provides the computational-level theory—this is *what* the cash register computes. Demonstrating *why* this is what the cash register computes involves

showing that our intuitive constraints on the cash register's operation map onto this computational-level theory. Note that for Marr the computational-level theory is a precise mathematical account of the function that the device computes. It is Marr's account of the computational level that we adopt here.

Why should a computational-level theory be defined in precise formal terms? Without a formal theory we must rely on incomplete or poorly specified intuitions that are not likely to result in a consistent computational-level theory. An inconsistent theory is, of course, valueless, because, from an inconsistency anything follows. Furthermore, appeal to mere intuition is ultimately circular, as the goal of a computational-level theory of reasoning is to explain our intuitions, and thus cannot simply take them for granted. Providing a consistent formal account of our intuitions in any domain is a difficult, but unavoidable, challenge.[3]

In the psychology of reasoning, demonstrating the appropriateness of computational-level theories has not been a prime concern. The standard approach has been to borrow *normative* theories, about what one should or should not do on a reasoning task, from logic and mathematics. However, the last 30 years of reasoning research has been notable, largely because of the mismatch observed between these normative accounts and subjects' behaviour. This indicates that these normative accounts can not provide appropriate computational-level theories of the tasks investigated by reasoning researchers. Normative theories and computational-level theories play different roles. Only the latter must be descriptively adequate to subjects' task performance. Although a normative theory may play an important role in inspiring the development of a reasoning task, it still remains an empirical question whether it provides a descriptively adequate computational-level theory.

To illustrate the difference between normative and computational-level accounts, let us consider an example. Suppose that you find some unknown device and wonder what its function might be. Perhaps, observing its behaviour, you suppose that it may be performing arithmetical calculations. To make this conjecture is to make a specific hypothesis about the computational-level theory appropriate for describing the device. On this assumption, you might give the device certain inputs, which you interpret as framing arithmetical problems. It may turn out, of course, that the outputs you receive do not appear to be interpretable as solutions to these, or perhaps any other, arithmetical problems. This may indicate that your computational-level theory is inappropriate, particularly if you cannot interpret most of the outputs as correct answers. You

[3]The problem of providing a consistent formalisation of intuitions in any domain is extremely difficult. Even providing a computational-level theory for a calculator has proved to be an enormous intellectual challenge. For example, Frege's (1950) formalisation of arithmetic succumbed to an unexpected paradox, due to Russell, which demonstrated the inconsistency of what appeared to be intuitively consistent intuitions (see Haack, 1978 for discussion).

may therefore search for an alternative computational-level explanation—perhaps the device is not doing arithmetic, but is solving differential equations. Thus, a computational-level theory must not only be normatively justified, it must also be descriptively adequate in a way that merely normative theories need not be. There is no doubt that arithmetic is a normative theory; what is in doubt is whether arithmetic is the appropriate normative theory to describe the behaviour of this device.

Similarly, in the psychology of reasoning, theorists cannot derive appropriate computational-level theories by reflecting on normative considerations alone, but only by attempting to use those theories to describe human inference. It is not controversial that logic provides a good normative theory of deductive inference—the question is: do people perform deductive inferences?.

The same point applies to tasks. In the case of the device, we may mistakenly interpret a set of inputs as posing an arithmetical problem, when the device consistently interprets these inputs as posing problems in solving differential equations. The experimenter cannot legislate concerning the nature of the task. Similarly, we may mistakenly interpret a psychological task as posing a deductive reasoning problem, when subjects consistently interpret the task as posing some other kind of well-defined problem. We suggest later on that Wason's selection task, for example, poses a problem of probabilistic optimal data selection, rather than a problem of logical inference, as is frequently assumed.

Only two of the four theories of reasoning that we discussed earlier—mental logics and mental models theory—embody a computational-level theory, in the sense that we have just described. The heuristic approach, as we mentioned earlier, does not by definition attempt to account for people's general inferential performance, but must supplement some theory that does provide such an account. Moreover, pragmatic reasoning schemas do not constitute a formal theory of reasoning. The schemata for deontic reasoning, for example, simply embody intuitions about appropriate deontic rules for use in specific situations. They do not have the goal of providing a computational-level theory of deontic or any other kind of reasoning.[4]

In contrast, mental logics and mental models embody logic as a computational-level theory. This is self-evident for the mental logic approach. But it also follows immediately for the mental models approach, given that the goal of mental models theory is to provide a mechanism for conducting logically valid deductive inference, as we observed earlier. Johnson-Laird and Byrne (1991) note that logic does not exhaust the computational-level of theory on the mental models account. Specifically, they outline three intuitive constraints on

[4]Proponents of pragmatic schemas are, however, concerned with computational-level issues. For example, Holyoak and Cheng (in press) describe considerations from jurisprudence which may serve as a starting point for a computational-level theory of aspects of deontic reasoning.

the kinds of deductive inferences that people actually draw. These constraints restrict human deductive inference to a subset of logically valid deductive inferences. But, we shall argue, logic is an inappropriate computational-level theory, not because it admits too many inferences, but because it admits too few: specifically, everyday defeasible inferences are not logically valid. Thus, in this context, we need not discuss Johnson-Laird and Byrne's (1991) additional computational-level constraints further.

Is Logic an Appropriate Computational-level Theory of Everyday Inference?

We saw earlier that applying logic to everyday defeasible reasoning requires a non-monotonic logic (e.g. Clark, 1978, McCarthy, 1980, McDermott & Doyle, 1980, Reiter, 1980, 1985, see also collection edited by Ginsberg, 1987) and that using such logics is computationally intractable. We now consider the prior question (see our Introduction) of whether non-monotonic logics can serve as adequate computational-level theories of human inference. Unfortunately non-monotonic logics also prove to be inadequate at this level of explanation (Harman, 1986; Israel, 1980; McDermott, 1987; Oaksford & Chater, 1991, 1992, 1993; Rips, 1994).

A crucial and ubiquitous problem for all these accounts arises when there is conflict between the conclusions drawn by different default rules. For example, suppose you are considering the following two default rules:

(1) If x is an academic & *there is no reason to suppose otherwise*, then x is unfit.
(2) If x is a runner & *there is no reason to suppose otherwise*, then x is fit.

and the fact that:

(3) Fred is both an academic and a runner.

Is Fred fit or unfit? The conclusion seems to depend on the order in which you apply the rules (see Oaksford & Chater, 1991, for a more formal example). Taking rule (1) first, because Fred is an academic and there is no reason to suppose otherwise—because, *crucially*, you have not yet considered rule (2)—you may infer that he is unfit—hence rule (2) is not now applicable. Taking rule (2) first, because Fred is a runner and there is no reason to suppose otherwise—because, *crucially*, you have not yet considered rule (1)—you may infer that he is fit—hence rule (1) is not now applicable. Because conclusions should be order-independent, the only possible conclusion is the wholly uninformative one that Fred is either fit or unfit. However, the intuitively obvious conclusion from this information is that Fred is fit—academics are typically unfit because they do not exercise, which does not apply to academic runners who clearly do exercise. The problem of conflicting defaults is widely recognised in AI as a central and

unsolved problem in knowledge representation (McDermott, 1987). Examples such as these indicate that non-monotonic logics are not appropriate computational-level theories of defeasible inference because they fail to capture peoples' intuitions about the appropriate inferences to draw.

If, as we have argued, everyday reasoning is non-logical, then mental logics and mental models would seem to be unable to generalise beyond the laboratory. But recent psychological results indicate that logic-based models may be inappropriate even *within* the laboratory. Work on conditional reasoning indicates that subjects interpret conditional sentences as default rules (Holland, Holyoak, Nisbett, & Thagard, 1986) even in laboratory tasks (Holyoak & Spellman, 1993; Oaksford et al., 1990). Byrne (1989), and Cummins, Lubart, Alksnis, and Rist (1991), have shown that background information derived from stored world knowledge can affect inferential performance (see also, Markovits, 1984, 1985). Specifically they showed that *additional antecedents* influence the inferences conditional statements allow. For example:

(1) If she has an essay to write then she will study late in the library.
(a) *Additional Antecedent:* The library is closed.

(1) could be used to predict that she will study late in the library if she has an essay to write. This is an inference by *modus ponens*. However, including information about an additional antecedent (a) *defeats* this inference (Byrne, 1989). Moreover, confidence in this inference reduces for rules that possess many alternative antecedents even when this information is only implicit (Cummins et al., 1991). Additional antecedents also affect inferences by *modus tollens*. If she does not study late in the library, you can infer that she didn't have an essay to write, unless the library was closed. Explicitly providing information about alternative antecedents defeats *modus tollens* (Byrne, 1989) and reduces confidence in rules that possess many alternative antecedents even when this information is only implicit (Cummins et al., 1991). In sum, people treat conditionals in laboratory reasoning tasks as default rules.

Advocates of mental logics and mental models, although aware of these arguments (Garnham, 1993; Johnson-Laird & Byrne, 1991; Rips, 1994), present a variety of proposals that may be thought to deflect these difficulties. We consider these later, and argue that they do not succeed.

Theories of Reasoning and the Computational Level

Mental Logics. Mental logicians appear to have dismissed the influence of default rules on reasoning as an interfering pragmatic or performance factor (Braine, Reiser, & Rumain, 1984; Rumain, Connell, & Braine, 1983). 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 driven to explore non-monotonic logics to capture just the phenomenon the mental logicians dismiss (see e.g. the collection edited by Ginsberg, 1987). As we have seen, embracing the defeasibility of everyday inference, these researchers immediately confront unsolved problems at both the algorithmic and the computational levels. Mental logic researchers, by contrast, have attempted to avoid these difficulties by maintaining—at least with respect to the experimental data they consider—that reasoning is in fact monotonic.

Perhaps the best worked out example is Politzer and Braine's (1991) attempt to deny that the data that we examined earlier from Byrne (1989) and Cummins et al. (1991) reflect defeasible inferential processes. We outline their position, and argue that it involves a fundamental misunderstanding of the nature of everyday, defeasible reasoning.

Politzer and Braine (1991) argue that Byrne's (1989) results do not show that additional information can defeat (or suppress) *modus ponens* because the premises result in an inconsistency.[5] Their argument is as follows. Byrne presented subjects with premises such as:

(3) If she has an essay to write then she will study late in the library
(4) She has an essay to write

in response to which subjects sponteneously make the inference by *modus ponens* that she will study late in the library. However, adding a further premise:

(5) If the library stays open then she will study late in the library

leads to a significant reduction in the number of subjects concluding that she will study late in the library. Subjects instead conclude that she may or may not study late in the library. Byrne (1989, p.76) describes this effect as showing "that context can suppress . . . valid . . . inferences." Politzer and Braine (1991) argue that general knowledge of libraries mean that (3)–(5) are likely to lead subjects to add:

(6) If she studies late in the library then necessarily the library stays open

to their premise set because (5) "actually expresses a necessary condition", i.e.

(5′) If the library is closed, then she cannot study late in the library.

[5]We here ignore Byrne's (1991) response to Politzer and Braine (1991) because we concur with O'Brien (1993) that Byrne misrepresents Politzer and Braine's argument.

But now there is an inconsistency because, (3) and (6) entail:

(7) If she has an essay to write then necessarily the library stays open

which subjects know to be false. Politzer and Braine argue that subjects therefore question the literal truth of (3) and hence fail to infer that she will study late. They also suggest that all putative cases of suppression of *modus ponens* are cases where one can question the literal truth of the premises.

Politzer and Braine's modal argument is not valid. But it is not necessary to delve into the technicalities (outlined in Appendix 2) to appreciate that this line of reasoning cannot be sound. First, intuitively (3)–(5) do not seem to be mutually inconsistent. And Politzer and Braine's argument that they are, given appropriate world knowledge, is not compelling. The crucial conclusion (6) is intuitively and logically bizarre: it suggests that a contingent truth about whether somebody studies late in the library implies that it could be a necessary truth that the library stays open. But whether or not somebody works late cannot make it necessary (in a logical, physical, causal, or any other substantive sense of necessity) that the library stays open, because counter-examples abound: she might break into the library, be locked in accidentally, may have a key, be a friend of the librarian, and so on. As we show in Appendix 2, our intuition that this inference—that supposedly demonstrates the inconsistency in (3)–(5)—is invalid, is supported by the fact that it is also invalid in modal logic. Given that (6) does not follow, even if we grant that people may infer (5′) from world knowledge, the rest of Politzer and Braine's (1991) argument collapses.

Treating these rules as default rules, however, leads to a far more natural interpretation of these experimental materials. The "inconsistent" conclusion that the library stays open if she has an essay to write only looks aberrant because Politzer and Braine explicitly add (6) and (7) as derived theorems. This presentation makes "the library stays open" seem like the consequence of a false rule (7). However, by treating (3) as a default rule, we can see "the library stays open" for what it is—a default assumption. Interpret (3) as previously:

(3′) If she has an essay to write & *there is no reason to suppose otherwise*, then she will study late in the library.

Given (4) the second conjunct must be satisfied. This involves checking whether she will not study late in the library can be proved from (3), (4), and (5′). Assuming forward and backward chaining (Rips, 1983, 1994), (5′) provides a match that yields "the library is closed" as a subgoal. This cannot be proved from (3), (4), and (5′). However, by the closed world assumption used by AI systems, as noted earlier, (Hogger, 1984) *not*(the library is closed), i.e. the

library is open, can be inferred.[6] Consequently, that she will not study late cannot be proved either, and hence it is safe to infer that she will study late in the library. Therefore (4) leads to the apparently undesirable assumption that the library is open. This assumption is innocuous, however. Informally, you infer she will study late in the library because (i) she has an essay to write and (ii) although you don't know whether the library is open or not, with no evidence to the contrary, you assume that it is. Subjects' willingness to endorse the conclusion that she has an essay to write is therefore dependent on their willingness to make this assumption and it is this assumption that experimenters manipulate in the task. Thus interpreting conditionals as default rules makes much better sense of the observed performance in conditional reasoning tasks than the attempt to maintain a logical interpretation.

Rips (1994, p.270) takes a rather different line to Politzer and Braine, conceding that "Defeasible inferences must be extremely common in everyday thinking, and any general theory in AI or psychology must accommodate them". But he argues that default reasoning arises in the context of inductive inference and (Rips, 1994, p.411) that although "Oaksford and Chater [1991] may be right that inductive inference will eventually be the downfall of these [classical logicist] approaches" this does not vitiate the mental logic approach. Rips (1994, p.411) argues that nondemonstrative belief fixation may come about "in other ways than making it the conclusion of an argument". In addition to these "other ways", Rips assumes that people have considerable resources for deductive reasoning, and argues for a particular account of these in terms of natural deduction.

If our arguments are correct, then this intermediate position is not tenable. The conclusion that people do not interpret natural language conditionals logically, but rather interpret them as default rules (Holyoak & Spellman, 1993; Oaksford & Chater, 1992, 1993) applies to almost any reasoning that mental logicians attempt to explain. For example, Rips offers the following example as a paradigmatic case of deductive inference:

(9) If Calvin deposits 50 cents, he'll get a coke.
Calvin deposits 50 cents
Therefore, Calvin will get a coke.

Rips treats this inference as deductive and hence *modus ponens* applies. However, in the light of previous discussion, the conditional premise is clearly about as good an example of a default rule as one could find. Calvin won't get the coke if the machine is broken, if the cokes have run out, if the power is turned off, and so on.

[6]We use an AI interpretation of defaults here for illustration only. As we noted earlier, such interpretations of default rules are not in general adequate.

It is possible to reply, as seems implicit in Politzer and Braine (1991) and Rips (1994), that such additional circumstances do not show that the first premise is defeasible (and therefore that some non-monotonic inference regime must be invoked), but simply show that it is false, according to the standard, non-defeasible, interpretation of the conditional. But if this is how people interpret conditionals, then the only conditionals that people believe true will be those that never admit of counter-examples. Because any everyday conditional, including (9), admits exceptions, then all such conditionals will be false. Clearly, people do not reject such conditionals, but freely assert them, argue about whether they are true, and use them to guide their behaviour. This makes perfect sense if people interpret conditionals as default rules; it makes no sense at all if they interpret conditionals logically.

In summary, mental logicians have on the whole attempted to marginalise defeasible reasoning. One argument is to deny (O'Brien, 1993; Politzer & Braine, 1991) that the empirical evidence supports the claim that people view the rules used in laboratory task as default rules (Holyoak & Spellman, 1993; Oaksford & Chater, 1992, 1993). We have shown that these arguments are not valid. However, even if they were valid, the mental logician would still have to account for the many clear cut cases of default inferences that occur in everyday life outside the laboratory. Rips (1994) attempts to avoid this problem by arguing that most default inferences are inductive and that such processes do not have to involve argument. However, we argue that even the paradigm examples that mental logicians do intend to explain are not logically valid, but involve defeasible inference. Given that standard logic cannot provide an appropriate computational-level model of defeasible, uncertain reasoning one might expect that the mental *logician* would therefore embrace non-standard, *non-monotonic logics*. However, they are rightly cautious—such logics fail to characterise the intuitively correct inferences, and hence could not provide an appropriate computational-level theory.

Mental Models. Proposals for incorporating default reasoning into mental models (Johnson-Laird & Byrne, 1991) rely on incorporating default assumptions into the initial mental model of a set of premises. Reasoners recruit these assumptions from prior world knowledge and may undo them in the process of changing mental models. Mental model theorists claim that they thereby avoid the problem of consistency checking, because there is no need to search for counter-examples to default assumptions. This proposal does not resolve the problem of default inference. A generalisable theory of reasoning must address the problem of *which* default assumption(s) to incorporate in an initial representation. For example, suppose I tell you that "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). So mental models theory confronts the problem of non-monotonic reasoning head on.

Mental model theorists may argue that the problem of searching for counter-examples for default assumptions is part of a theory of memory retrieval that mental models, as a theory of inference, need not provide. Three arguments seem to vitiate this suggestion. First, AI treats these memory retrieval processes as *inferential* processes that a theory of inference should explain. Second, these memory retrieval processes involve the search for counter-examples. Therefore *in its own terms* these processes are exactly the type 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 these processes to explain the results of reasoning tasks.

In recent accounts (e.g. Johnson-Laird & Byrne, 1991) the explanation of various phenomena depends on the way in which an initial mental model of the premises is "fleshed-out". "Fleshing-out", for example, determines whether a disjunction is interpreted as exclusive or inclusive *or* (Johnson-Laird & Byrne, 1991, p.45); whether a conditional is interpreted as material implication or equivalence (Johnson-Laird & Byrne, 1991, pp.48–50) which in turn determines whether inferences by *modus tollens* will be performed; whether non-standard interpretations of the conditional are adopted (Johnson-Laird & Byrne, 1991, p.67), including content effects whereby the relation between antecedent and consequent affects the interpretation (Johnson-Laird & Byrne, 1991, pp.72–73); confirmation bias in Wason's selection task (Johnson-Laird & Byrne, 1991, p.80); and the search for counter-examples in syllogistic reasoning (Johnson-Laird & Byrne, 1991, p.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. Consequently it is reasonable to expect mental models theory to provide an account of how people retrieve relevant defaults from world knowledge. Appeal to fleshing-out is thus simply an *appeal* to a solution to the problem of everyday defeasible reasoning, it does not provide such a solution.

Garnham (1993) has suggested a related, but distinct, line of argument suggesting that mental models are applicable to non-monotonic reasoning, if there are restrictions on which models reasoners entertain. In particular, in non-monotonic reasoning Garnham proposes that people do not exhaustively search all possible models, but entertain only the most plausible models, perhaps even just the single most plausible model. In response to Garnham, the following default inference is considered by Chater & Oaksford (1993): if Fred eats a banana he peels it first, and Fred eats a banana, to the conclusion that he peeled it. As this inference is non-monotonic, there are many models in which the premise is true and the conclusion false—a friend may have peeled the banana,

Fred may have eaten it whole and so on. However, these models are not, in the absence of additional information, plausible. Much more plausible is the model in which Fred peeled the banana himself. To reason successfully about these matters, reasoners should consider only the plausible models.

Chater and Oaksford (1993) suggest that this line of reasoning has, in Russell's phrase, all the virtues of theft over honest toil. Mental models must assume as given a mechanism that distinguishes plausible from implausible models—and furthermore comes up with the most plausible models spontaneously. In other words, it presupposes a mechanism that is able to carry out inference to the best explanation—to devise and assess the plausibility of hypotheses to explain and be explained by known information. But inference to the best explanation is simply a paradigm case of non-monotonic inference. A mental models account in which the ability to construct just the right model (the best explanation) is a primitive operation finesses, rather than addresses, the problem of non-monotonic reasoning.

Mental models therefore inevitably seem to founder on either of two difficulties (Chater & Oaksford, 1993). Without some notion of which models are plausible and which are not, it will invariably be possible to construct some (implausible) model, even for the most persuasive of common-sense inferences, and hence mental models will license no common-sense inferences at all. On the other hand, if mental modellers presuppose some notion of plausibility, then they are simply assuming a solution to the problem of accounting for common-sense reasoning rather than explaining it.

Garnham's (1993) specific proposals for a theory of non-monotonic reasoning based on mental models take the latter course. Garnham argues that certain quite unexpected considerations may be sufficient to pick out which models people *should* take to be plausible (Garnham, 1993, p.63):

> The *should* is more likely to be cashed out in terms of what people can be expected to do, given their cognitive capacities, in particular the processing and capacity limitations of short-term memory working memory and the organisation and retrieval of information from long-term memory. Thus, people should consider revisions of their mental models that are required by a specific piece of information that has entered working memory, from long-term memory or elsewhere.

This does not, however, seem to help advocates of mental models. No doubt the organisation of human memory plays an important role in human reasoning; perhaps memory is organised to allow easy access to plausible models, and restricted access to implausible models; perhaps relevant information feeds into a short-term store as required, and irrelevant information is suppressed, and so on. This is just to say that the organisation of human memory profoundly affects human common-sense reasoning processes. This is a view, as we noted earlier,

with which most theorists would probably concur. However, it goes no way at all to providing an account of how such reasoning occurs. Moreover, it does not indicate why such an account should look like, or have any place for, mental models theory.

Apart from appealing to memory, Garnham (1993) also suggests that simple strategies can guide the model-building process. So, for example (Garnham 1993, p.63) " . . . revisions that falsify a conclusion consistent with the current model should not be considered, unless they are unavoidable" and "A conclusion can be accepted (tentatively, if it is defeasible) if there is some model of the premises that will accommodate it." However, Chater and Oaksford (1993) argue that these proposals cannot distinguish good from bad inferences, without covert assumptions concerning which models are plausible. With regard to Garnham's first principle, suppose you learn that Fred ate a banana, and create a model in which he peeled the banana before eating it. Suppose you then discover that Fred choked on the banana skin. Your natural reaction would be to overturn the tentative conclusion that Fred peeled the banana before eating it, and infer instead that Fred attempted the whole banana, peel and all. This seems more plausible than alternative models, where Fred peels and eats his banana and then eats the skin too, and so on. However, Garnham's principle does not allow such a retraction to occur, as revision of the tentative conclusion is certainly not unavoidable—just rather unlikely. Unless there is some hidden appeal to plausibility, and hence to a prior solution to the problem of non-monotonic inference, Garnham's principle will not allow us to account for this obvious everyday inference.

Let us turn to Garnham's second principle, that reasoners can accept (albeit tentatively) any proposition that some model of the premises can accommodate. Chater and Oaksford (1993) argue that this principle seems to lead immediately to inferential anarchy. For example, there is a model in which Fred eats a banana and a pig is sitting on the roof of his house (assuming no information to the contrary). Thus Garnham's second principle licenses this bizarre conclusion, which reasoners tentatively accept. Of course, similar reasoning can also lead to the acceptance of the opposite conclusion (although, by the first principle, the first of these to be accepted will preclude the other from being accepted). There is, of course, a very large difference between models in which there is and is not a pig on the roof—the former will generally be less plausible. But plausibility is what is to be explained, and thus cannot itself be presupposed in explanation.

In summary, as might be expected from its reliance on logic as a computational-level theory, mental models theory fares no better than mental logic in dealing with defeasible everyday reasoning.

Probabilistic Approaches

We have argued that those theories of reasoning that have a computational-level account employ logic in that role, and hence cannot generalise to everyday

inference. We have, furthermore, argued that these theories provide inappropriate accounts of much reasoning in the laboratory. The problem for logical approaches is that real human reasoning is uncertain. So rather than attempting to apply deductive logic, the calculus of *certain* reasoning, to model uncertainty, why not apply probability theory, i.e. the calculus of *uncertain* reasoning? In this section, we illustrate this approach using Wason's selection task. This task has been of central importance to the development of the psychology of reasoning, and the difficulty of reconciling subjects' performance with logic has even been taken to question human rationality (Stich, 1985, 1990).

Wason's Selection Task. Wason's (1966, 1968) task requires subjects to assess whether some evidence is relevant to the truth or falsity of a conditional rule of the form *if p then q*, where by convention "*p*" stands for the antecedent clause of the conditional and "*q*" for the consequent clause. The task involves four cards each having a number on one side and a letter on the other, and a rule, e.g. *if there is a vowel on one side (p), then there is an even number on the other side (q)*. The four cards show an "A" (*p* card), a "K" (*not-p* card), a "2" (*q* card), and a "7" (*not-q* card). Subjects select those cards they must turn over to determine whether the rule is true or false. Typical results were: *p* and *q* cards (46%); *p* card only (33%) *p, q* and *not-q* cards (7%); *p* and *not-q* cards (4%) (Johnson-Laird & Wason, 1970).

Logic, and Popper's (1959) account of falsification, provide the standard computational-level theory of the selection task. Popper argued that observation cannot prove the truth of a scientific law because it is always possible that the next instance of the law observed will be falsifying. However, you can be logically certain that a law is false by uncovering a single counter-example. Popper's account means that scientific reasoning is fundamentally deductive in character—scientists must establish a logical contradiction between putative laws and observation. Hence looking for false (*p* and *not-q*) instances should be the goal of scientific inquiry. However, in the selection task subjects typically select cards that could *confirm* the rule, i.e. the *p* and *q* cards. Thus, it appears that the logical computational-level theory based on Popper, is descriptively inadequate, and hence inappropriate as a computational-level theory.[7]

It is a common assumption that Wason's selection task is a deductive task. However, as we noted earlier, this assumption is called into question if we can show that an alternative computational-level theory of the task is more descriptively adequate. We have recently recast Wason's selection task

[7]It is an independent and controversial issue whether Popper's account is normatively justified. Few modern philosophers of science endorse the falsificationist position; and there has been renewed interest in probabilistic models of scientific inference (Earman, 1992; Horwich, 1982; Howson & Urbach, 1989).

probabilistically, as a problem of Bayesian optimal data selection (Oaksford & Chater, 1994). Any problem of deciding what experiment to perform next, or which observation is worth making, is a problem of optimal data selection. Suppose that you are testing the hypothesis that eating tripe makes people feel sick. In collecting evidence, should you ask known tripe-eaters or tripe-avoiders whether they feel sick? Should you ask people known to be sick, or known not to be, whether they have eaten tripe? This case is analogous to the selection task. Logically, you can write the hypothesis as a conditional sentence, if you eat tripe (p) then you feel sick (q). The groups of people that you may investigate then correspond to the various visible card options, p, *not-p*, q, and *not-q*. In practice, who is available will influence decisions about who to investigate. The selection task abstracts away from this practical detail by presenting one example of each potential source of data. In terms of our everyday example, it is like coming across four people, one known to have eaten tripe, one known not to have eaten tripe, one known to feel sick, and one known not to feel sick. You must then judge which of these people you should question about how they feel or what they have eaten.

Oaksford and Chater (1994) suggest that hypothesis-testers should select data points expected to provide the greatest information gain in deciding between two hypotheses: (i) that the task rule, if p then q, is true, i.e. ps are invariably associated with qs, and (ii) that the occurrence of ps and qs are independent. For each hypothesis, Oaksford and Chater (1994) define a probability model that derives from the prior probability of each hypothesis (which for most purposes they assume to be equally likely, i.e. both = 0.5), and the probabilities of p and of q in the task rule. They define information gain as the difference between the uncertainty *before* receiving some data and the uncertainty *after* receiving that data, where they measure uncertainty using Shannon-Wiener information. Thus Oaksford and Chater define the information gain of a data point D as:

$$\text{Information before receiving } D\text{: } I(H_i) = -\sum_{i=1}^{n} P(H_i)\log_2 P(H_i)$$

$$\text{Information after receiving } D\text{: } I(H_i|D) = -\sum_{i=1}^{n} P(H_i|D)\log_2 P(H_i|D)$$

$$\text{Information gain: } I_g = I(H_i) - I(H_i|D)$$

Oaksford and Chater calculate the $P(H_i/D)$ terms using Bayes' theorem. Thus information gain is the difference between the information contained in the *prior* probability of a hypothesis (H_i) and the information contained in the *posterior* probability of that hypothesis given some data D.

In the selection task, however, when choosing which data point to examine further (that is, which card to turn), the subject does not know what the data point will be (that is, what will be the value of the hidden face). So they could

not calculate actual information gain. However, subjects can compute *expected* information gain. Expected information gain is calculated with respect to all possible data outcomes, e.g. for the *p* card, *q*, and *not-q*.

To model selection-task data, Oaksford and Chater (1994) calculated the expected information gain of each card assuming that the properties described in *p* and *q* are rare. Klayman and Ha (1987) make a similar assumption in accounting for related data on Wason's (1960) 2-4-6 task. The order in expected information gain is:

$$E(I_g(p)) > E(I_g(q)) > E(I_g(not\text{-}q)) > E(I_g(not\text{-}p))$$

This corresponds to the observed frequency of card selections in Wason's task: $p > q > not\text{-}q > not\text{-}p$ and thus explains the predominance of *p* and *q* card selections as a rational inductive strategy. Using this style of explanation, Oaksford and Chater (1994) model a wide range of data concerning: the non-independence of card selections (Pollard, 1985); the negations paradigm (e.g. Evans & Lynch, 1973); the therapy experiments (e.g. Wason, 1969); the reduced array selection task (Johnson-Laird & Wason, 1970); work on so-called fictional outcomes (Kirby, 1994); and deontic versions of the selection task, including perspective and rule-type manipulations (e.g. Cheng & Holyoak, 1985); and the manipulation of probabilities and utilities in deontic tasks (Kirby, 1994).

In short, using the calculus of uncertainty, we have provided a new computational-level theory of the selection task, which appears to be descriptively adequate for a wide range of experimental data. The possibility of probabilistic models of the selection task has been raised, but not pursued, in the past (Fischhoff & Beyth-Marom, 1983; Klayman & Ha, 1987; Rips, 1990).[8] Furthermore, there have been some informal (Manktelow & Over, 1991) and formal (Kirby, 1994) probabilistic treatments of certain aspects of the selection task. We believe that we can extend our probabilistic computational-level theory of the selection task to other inductive reasoning tasks. In these areas, a number of probabilistic computational-level accounts have demonstrated descriptive adequacy on tasks that, like the selection task, were previously thought to impugn human rationality (Anderson, 1990, 1991a, 1991b; Birnbaum, 1983; Cheng & Novick, 1990, 1991, 1992; Gigerenzer, Hell, & Black, 1988; Gigerenzer & Hoffrage, in press; Gigerenzer, Hoffrage, & Kleinbölting, 1991; Gigerenzer & Murray, 1987).

[8]It is interesting that mental logicians such as Rips (1990) have advocated probabilistic models of the selection task, suggesting that they too agree that logic does not provide a descriptively adequate computational-level theory for this task.

CONCLUSIONS

We have argued that current theories of reasoning fail to generalise to defeasible everyday inference. They are inadequate at both Marr's algorithmic and computational levels because they are unable either to provide tractable algorithms for defeasible inference, or to provide a computational-level theory that characterises the inferences that people draw. We focused on the computational level, arguing that current reasoning theories use logic as a computational-level theory, where such a theory is evident at all. We suggested that probabilistic computational-level accounts are more appropriate for capturing the uncertain character of human reasoning and illustrated this approach for Wason's selection task.

Our approach is consistent with similar proposals of Anderson (1990) and Evans (1993). Anderson argues for what he calls "rational analyses" of a wide range of cognitive phenomena, though not focusing on traditional reasoning tasks. A rational analysis provides a computational-level explanation of a task, by showing that behaviour is optimally adapted to the task environment. In such analyses, the task environment is usually characterised probabilistically. Our account of Wason's selection task thus provides a rational analysis in Anderson's sense. We suggest that it may be fruitful for reasoning theory to adopt Anderson's programme for other tasks. This also seems consistent with Evans' recent distinction between rationality$_1$ and rationality$_2$ theories of reasoning (Evans, 1993; Evans, Over, & Manktelow, 1993). A rationality$_1$ theory aims to explain how behaviour is suitable for fulfilling the organism's goals; thus it provides a computational-level theory or rational analysis. A rationality$_2$ theory concerns the processes from which behaviour arises, and hence focuses on the algorithmic level.

Marr argued that the computational level of explanation is primary. We have suggested a fundamental shift, from logic to probability, in the framework in which to construct computational-level theories of reasoning. But the algorithmic problem remains. We have already noted that probabilistic inference, like logical inference, is, in full generality, computationally intractable. What is encouraging is that probabilistic computational-level theories promise to explain much data on human reasoning, without having to engage so-far-unsolved questions concerning the algorithms that implement world knowledge in the brain. A complete psychological theory must, of course, provide explanations at each of Marr's three levels. But as Marr argued, only with a descriptively adequate computational-level theory in hand is it possible sensibly to ask which algorithms carry out these computations, and how these algorithms are implemented in the brain.

Manuscript received 11 April 1994
Revised manuscript received 25 January 1995

REFERENCES

Anderson, J.R. (1990). *The adaptive character of thought*. Hillsdale, NJ: Lawrence Erlbaum Associates Inc.

Anderson, J.R. (1991a). Is human cognition adaptive? *Behavioral and Brain Sciences*, *14*, 471–517.

Anderson, J.R. (1991b). The adaptive nature of human categorization. *Psychological Review*, *98*, 409–429.

Birnbaum, M.H. (1983). Base rates in Bayesian inference: Signal detection analysis of the cab problem. *American Journal of Psychology*, *96*, 85–94.

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., & 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.

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

Byrne, R.M.J. (1991). Can valid inferences be suppressed. *Cognition*, *39*, 71–78.

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

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

Chater, N., & Oaksford, M. (1993). Logicism, mental models and everyday reasoning: Reply to Garnham. *Mind & Language*, *8*, 72–89.

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

Cheng, P.W., & Holyoak, K.J. (1989). On the natural selection of reasoning theories.*Cognition*, *33*, 285–313.

Cheng, P.W., & Novick, L.R. (1990). A probabilistic contrast model of causal induction. *Journal of Personality and Social Psychology*, *58*, 545–567.

Cheng, P.W., & Novick, L.R. (1991). Causes versus enabling conditions. *Cognition*, *58*, 83–120.

Cheng, P.W., & Novick, L.R. (1992). Covariation in natural causal induction. *Psychological Review*, *99*, 365–382.

Clark, H.H. (1977). Bridging. In P.N. Johnson-Laird, & P.C. Wason (Eds.), *Thinking* (pp. 411–420). Cambridge: Cambridge University Press.

Clark, K.L. (1978). Negation as failure. In H. Gallaire, & J. Minker (Eds.), *Logic and databases* (pp. 293–322). New York: Plenum Press.

Cook, S. (1971). The complexity of theorem proving procedures. In *The third annual symposium on the theory of computing* (pp. 151–158). New York, NY.

Cosmides, L. 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., & Rist, R. (1991). Conditional reasoning and causation. *Memory & Cognition*, *19*, 274–282.

Earman, J. (1992). *Bayes or bust?* Cambridge, MA: MIT Press.

Enderton, H.B. (1972). *A mathematical introduction to logic*. New York: Academic Press.

Evans, J.St.B.T. (1983). Selective processes in reasoning. In J.St.B.T. Evans, (Ed.), *Thinking and reasoning: Psychological approaches*. London: Routledge & Kegan Paul.

Evans, J.St.B.T. (1984). Heuristic and analytic processes in reasoning. *British Journal of Psychology*, *75*, 451–468.

Evans, J.St.B.T. (1989). *Bias in human reasoning: Causes and consequences*. London: Lawrence Erlbaum Associates Ltd.

Evans, J.St.B.T. (1991). Theories of human reasoning: The fragmented state of the art. *Theory & Psychology*, *1*, 83–105.

Evans, J.St.B.T. (1993). Bias and rationality. In K.I. Manktelow, & D.E. Over (Eds.), *Rationality* (pp. 6–30). London: Routledge.

Evans, J.St.B.T. (in press). Relevance and reasoning. In S.E. Newstead, & J.St.B.T. Evans (Eds.), *Current directions in thinking and reasoning*. Hove, UK: Lawrence Erlbaum Associates Ltd.

Evans, J.St.B.T., & Lynch, J.S. (1973). Matching bias in the selection task. *British Journal of Psychology, 64*, 391–397.

Evans, J.St.B.T., Over, D.E., & Manktelow, K.I. (1993). Reasoning, decision making and rationality. *Cognition, 49*, 165–187.

Fischhoff, B., & Beyth-Marom, R. (1983). Hypothesis evaluation from a Bayesian perspective. *Psychological Review, 90*, 239–260.

Fodor, J.A. (1983). *Modularity of mind*. Cambridge MA: MIT Press.

Frege, G. (1950). *The foundations of arithmetic* (Translated by J.L. Austin). Oxford: Basil Blackwell (Originally published 1884).

Garey, M.R. & Johnson, D.S. (1979). *Computers and intractability: A guide to the theory of NP-completeness*. San Francisco: W.H. Freeman.

Garnham, A. (1993). Is logicist cognitive science possible? *Mind & Language, 8*, 49–71.

Garrod, S., & Sanford, A.J. (1977). Interpreting anaphoric relations: The integration of semantic information while reading. *Journal of Verbal Learning and Verbal Behavior, 16*, 77–90.

Gentzen, G. (1934). Untersuchungen über das logische Schliessen. *Mathematische Zeitschrifft, 39*, 176–210.

Gigerenzer, G., Hell, W., & Blank, H. (1988). Presentation and content: The use of base rates as a continuous variable. *Journal of Experimental Psychology: Human Perception and Performance, 14*, 513–525.

Gigerenzer, G., & Hoffrage, U. (in press). How to improve Bayesian reasoning without instruction: Frequency domains.

Gigerenzer, G., Hoffrage, U., & Kleinbölting, H. (1991). Probabilistic mental models: A Brunswickian theory of confidence. *Psychological Review, 98*, 506–528.

Gigerenzer, G., & Murray, D.J. (1987). *Cognition as intuitive statistics*. Hillsdale, NJ: Lawrence Erlbaum Associates Inc.

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* (pp. 65–76). Norwood, NJ: Ablex.

Haack, S. (1978). *Philosophy of logics*. Cambridge: Cambridge University Press.

Harman, G. (1986). *Change in view*. Cambridge, MA: MIT Press.

Hayes, P. (1978). The naive physics manifesto. In D. Michie (Ed.), *Expert systems in the microelectronic age*. Edinburgh, UK: Edinburgh University Press.

Hayes, P. (1979). The logic of frames. In D. Metzing (Ed.), *Frame conceptions and text understanding* (pp. 40–61). Berlin: Walter de Gruyter & Co.

Hayes, P. (1984a). The second naive physics manifesto. In J. Hobbs (Ed.), *Formal theories of the commonsense world*. Hillsdale, NJ: Ablex.

Hayes, P. (1984b). Liquids. In J. Hobbs (Ed.), *Formal theories of the commonsense world*. Hillsdale, NJ: Ablex.

Henle, M. (1962). On the relation between logic and thinking. *Psychological Review, 69*, 366–378.

Hogger, C.J. (1984). *An introduction to logic programming*. New York: Academic Press.

Holland, J.H., Holyoak, K.J. Nisbett, R.E., & Thagard, P.R. (1986). *Induction: Processes of inference, learning and discovery*. Cambridge, MA: MIT Press.

Holyoak, K.J., & Cheng, P.W. (in press). Pragmatic reasoning with a point of view. *Thinking & Reasoning*.

Holyoak, K.J., & Spellman, B.A. (1993). Thinking. *Annual Review of Psychology, 44*, 265–315.

Horowitz, E., & Sahni, S. (1978). *Fundamentals of computer algorithms*, Rockville, Maryland:

Computer Science Press, Inc.
Horwich, P. (1982). *Probability and evidence*. Cambridge Univeristy Press.
Howsen, C., & Urbach, P. (1989). *Scientific reasoning: The Bayesian approach*. La Salle, Illinois: Open Court.
Hughes, G.E., & Cresswell, M.J. (1968). *An introduction to modal logic*. London: Methuen & Co. Ltd.
Inhelder, B. & Piaget, J. (1958). *The growth of logical reasoning*. New York: Basic Books.
Israel, D.J. (1980). What's wrong with nonmonotonic logic? In *Proceedings of AAAI-80*, (pp. 99–101).
Johnson-Laird, P.N. (1975). Models of deduction. In R.J. Falmagne (Ed.), *Reasoning: Representation and process*. Hillsdale, NJ: Lawrence Erlbaum Associates Inc.
Johnson-Laird, P.N. (1983). *Mental models*. Cambridge: Cambridge University Press.
Johnson-Laird, P.N., & Byrne, R.M.J. (1991). *Deduction*. Hillsdale, NJ: Lawrence Erlbaum Associates Inc.
Johnson-Laird, P.N., & Wason, P.C. (1970). A theoretical analysis of insight into a reasoning task. *Cognitive Psychology*, *1*, 134–148.
Kintsch, W., & Van Dijk, T.A. (1978). Toward a model of text comprehension and production. *Psychological Review*, *85*, 363–394.
Kirby, K.N. (1994). Probabilities and utilities of fictional outcomes in Wason's four-card selection task. *Cognition*, *51*, 1–28.
Klayman, J., & Ha, Y. (1987). Confirmation, disconfirmation and information in hypothesis testing. *Psychological Review*, *94*, 211–228.
Manktelow, K.I., & Over, D.E. (1991). Social roles and utilities in reasoning with deontic conditionals. *Cognition*, *39*, 85–105.
Markovits, H. (1984). Awareness of the "possible" as a mediator of formal thinking in conditional reasoning problems. *British Journal of Psychology*, *75*, 367–376.
Markovits, H. (1985). Incorrect conditional reasoning among adults: Competence or performance. *British Journal of Psychology*, *76*, 241–247.
Marr, D. (1982). *Vision*. San Francisco, CA: W.H. Freeman & Co.
McArthur, D.J. (1982). Computer vision and perceptual psychology. *Psychological Bulletin*, *92*, 283–309.
McCarthy, J.M. (1980). Circumscription—a form of non–monotonic reasoning. *Artificial Intelligence*, *13*, 27–39.
McCarthy, J.M., & Hayes, P. (1969). Some philosophical problems from the standpoint of Artificial Intelligence. In B. Meltzer & D. Michie (Eds.), *Machine intelligence*, *4*. New York: Elsevier.
McDermott, D. (1987). A critique of pure reason. *Computational Intelligence*, *3*, 151–160.
McDermott, D., & Doyle, J. (1980). Non–monotonic logic I. *Artifical Intelligence*, *13*, 41–72.
Minsky, M. (1975). Frame-system theory. In R. Schank & B.L. Nash-Webber (Eds.), *Theoretical issues in natural language processing*, Cambridge, MA, June 10–13, 1975.
Newell, A., & Simon, H.A. (1972). *Human problem solving*. Englewood Cliff, NJ: Prentice-Hall.
Oaksford, M., & Chater, N. (1991). Against logicist cognitive science. *Mind & Language*, *6*, 1–38.
Oaksford, M., & Chater, N. (1992). Bounded rationality in taking risks and drawing inferences. *Theory & Psychology*, *2*, 225–230.
Oaksford, M., & Chater, N. (1993). Reasoning theories and bounded rationality. In K.I. Manktelow, & D.E. Over (Eds.), *Rationality* (pp. 31–60). London: Routledge.
Oaksford, M., & Chater, N. (1994). A rational analysis of the selection task as optimal data selection. *Psychological Review*, *101*, 608–631.
Oaksford, M., Chater, N., & Stenning, K. (1990). Connectionism, classical cognitive science and experimental psychology. *AI & Society*, *4*, 73–90. Also in A. Clark & R. Lutz (Eds.). (1992). *Connectionism in context* (pp. 57–74). Berlin: Springer-Verlag.

O'Brien, D.P. (1993). Mental logic and human irrationality. In K.I. Manktelow, & D.E. Over (Eds.), *Rationality* (pp. 110–135). London: Routledge.

O'Brien, E.J., Shank, D.M., Myers, J.L., & Rayner, K. (1988). Elaborative inferences during reading: Do they occur on line? *Journal of Experimental Psychology: Learning, Memory & Cognition, 14*, 410–420.

Osherson, D. (1975). Logic and models of logical thinking. In R.J. Falmagne (Ed.), *Reasoning: Representation and process*. Hillsdale, NJ: Lawrence Erlbaum Associates Inc.

Politzer, G., & Braine, M.D.S. (1991). Responses to inconsistent premises cannot count as suppression of valid inferences. *Cognition, 38*, 103–108.

Pollard, P. (1985). Nonindependence of selections on the Wason selection task. *Bulletin of the Psychonomic Society, 23*, 317–320.

Popper, K.R. (1959). *The logic of scientific discovery*. London: Hutchinson.

Posner, M.I. (1978). *Chronometric investigations of the mind*. New York: Lawrence Erlbaum Associates Inc.

Pylshyn, Z.W. (Ed.) (1987). *The robot's dilemma: The frame problem in Artificial Intelligence*. Norwood, NJ: Ablex.

Reiter, R. (1985). A logic for default reasoning. *Artificial Intelligence, 13*, 81–132.

Reiter, R. (1985). On reasoning by default. In R. Brachman & H. Levesque (Eds.), *Readings in knowledge representation*. Los Altos, CA: Morgan Kaufman (originally published 1978).

Rips, L.J. (1983). Cognitive processes in propositional reasoning. *Psychological Review, 90*, 38–71.

Rips, L.J. (1990). Reasoning. *Annual Review of Psychology, 41*, 321–353.

Rips, L.J. (1994). *The psychology of proof*. Cambridge, MA: MIT Press.

Rumain, B., Connell, J., & 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*, 471–481.

Rumelhart, D.E. (1980). Schemata: The building blocks of cognition. In R.J. Spiro, B.C. Bruce, & W.F. Brewer (Eds.), *Theoretical issues in reading comprehension*. Hillsdale, NJ: Lawrence Erlbaum Associates Inc.

Schank, R.C. (1982). *Dynamic memory*. Cambridge: Cambridge University Press.

Sperber, D., Cara, F., & Girotto, V. (in press). Relevance explains the selection task. *Cognition*.

Stich, S. (1985). Could man be an irrational animal? *Synthese, 64*, 115–135.

Stich, S. (1990). *The fragmentation of reason*. Cambridge, MA: MIT Press.

Tsotsos, J.K. (1990). Analyzing vision at the complexity level. *Behavioral & Brain Sciences, 13*, 423–469.

Wason, P.C. (1960). On the failure to eliminate hypotheses in a conceptual task. *Quarterly Journal of Experimental Psychology, 12*, 129–140.

Wason, P.C. (1966). Reasoning. In B. Foss (Ed.), *New horizons in psychology*. Harmondsworth, UK: Penguin.

Wason, P.C. (1968). Reasoning about a rule. *Quarterly Journal of Experimental Psychology, 20*, 273–281.

Wason, P.C. (1969). Regression in reasoning. *British Journal of Psychology, 60*, 471–480.

APPENDIX 1: COMPUTATIONAL COMPLEXITY THEORY

Complexity theory derives a function describing the rate at which an algorithm consumes computational resources dependent on the size of the input, *n* (Garey & Johnson, 1979; Horowitz & Sahni, 1978). The crucial aspect of this function is its *order of magnitude*, O(), that reflects the rate at which resource demands increase with *n*:

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

For example, O(1) indicates that the number of times the algorithm executes basic machine operations 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 algorithm executes basic machine operations is some polynomial function of the input length; such algorithms are *polynomial time computable* (this class includes all algorithms of order lower than some polynomial function).

Complexity theory draws an important distinction between polynomial-time computable algorithms [$O(n^i)$ for some n], and *exponential-time* algorithms [for example, $O(2^n)$ or worse]. As *n* increases, exponential-time algorithms consume vastly greater resources than polynomial-time algorithms. This distinction marks the boundary between tractable (polynomial-time) and intractable (exponential-time) algorithms. Applying these distinctions to problems, a problem is polynomial-time computable if it has a polynomial-time algorithm. If all algorithms for the problem are exponential-time, then the problem is "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 that only possess polynomial-time algorithms that are non-deterministic are "in NP". NP-complete problems form a subclass of *NP-hard* problems. A problem is NP-hard if satisfiability reduces to it (Cook, 1971). A problem is NP-complete if it is NP-hard *and* is in NP. The class of NP-complete problems includes such classic families of problems as the "travelling salesman" problems. Whether any NP-complete problem is polynomial-time computable is unknown, but 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. In practice, computer scientists take the discovery that a problem is NP-complete to rule out the possibility of a real-time tractable implementation. In practical terms this may mean that for some *n* an algorithm that is NP-complete may not provide an answer in our lifetimes, if at all.

APPENDIX 2: THE VALIDITY OF POLITZER AND BRAINE'S (1991) MODAL ARGUMENT

We show that Politzer and Braine's argument is not valid and that it relies on inappropriately mixing modal and classical arguments. Politzer and Braine argue that (6) and (3) lead to (7) and that (6) is a necessary truth. On closer examination neither claim is sustainable. We note first that (6) does not follow from (5′), although a similar modal conclusion to (6) does follow on the assumption that the conditional in (5′) is interpreted as strict implication (*p could* not be true and *q* false) rather than the material conditional (*p is* not true and *q* false) (Haack, 1978). (8) follows from *not-q* $\prec$ *not-p* (where "$\prec$" is strict implication and where "*L*" is necessarily):

(8) $L(p \supset q)$

which means (6) should read:

(6′) Necessarily, if she studies late in the library, then the library stays open.

This inference is valid in Brouwer's System T, and systems S4 and S5, which form the basis of most modal logics (Hughes & Cresswell, 1968). In none of these systems, or, to our knowledge, in any modal logic, is the inference that Politzer and Braine's argument relies on [(*not-q* $\supset$ *not-p*) $\models$ (*p* $\supset$ *Lq*)], a valid inference. As Hughes and Cresswell (1968, p.27 footnote) observe, *p* $\supset$ *Lq* (7) is "often confused [with (8)] in ordinary discourse, sometimes with disastrous results." The result here is that (3) and (6′) do not entail (7), because (8) is equivalent to *Lp* $\supset$ *Lq*, but (3) and (4) do not lead to the conclusion that *necessarily* she will study late in the library. So, (3) and (6′) could not transitively entail (7). Consequently Politzer and Braine's argument is not valid. Moreover, far from being a necessary truth (6′) is strictly false, as it is possible that she studies late in the library while the library remains shut—she could break in, get accidentally locked in, and so on. Thus neither (6) nor (6′) express necessary truths as Politzer and Braine assert.