Mind & Language
Vol. 6 No. 1 Spring 1991

ISSN 0268–1064

Article

Against Logicist Cognitive Science

MIKE OAKSFORD* AND NICK CHATER

> It would not be unreasonable to describe Classical Cognitive Science as an extended attempt to apply the methods of proof theory to the modelling of thought. (Fodor & Pylyshyn, 1988, pp. 29–30)

1. Introduction

In this paper, we shall argue that the plausibility of classical, logicist cognitive science depends on its ability to provide a proof-theoretic account of the defeasible inferencing which is implicated in almost every area of cognitive activity. We shall show that such an account is unlikely to be forthcoming and hence cognition cannot be seen as mechanised proof theory.

* The order of authorship is arbitrary. We would like to thank Patrick Blackburn, Gordon D. A. Brown, Brian Butterworth, Robin Cooper, Nick Ellis, David Green, Nigel Harvey, Geoffrey Hunter, A. R. Jonckheere, Mike Malloch, Ken Manktelow, Marc Moens, David Over, Ullin Place, Barry Richards, Jerry Seligman, Neil Smith, Keith Stenning and two anonymous reviewers for helpful comments on earlier versions of this work. We should also like to thank the Knowledge and Inference Group at the Centre for Cognitive Science, University of Edinburgh, where we first discussed many of the ideas in this paper. Much of this work was carried out while Mike Oaksford was funded under ESRC research contract R000231282, and Nick Chater was funded by ESRC grant C00428622023. Correspondence regarding this article should be sent either to Mike Oaksford, Cognitive Neurocomputation Unit, Department of Psychology, University of Wales, Bangor, Gwynedd, LL57 2DG, Wales, U.K. or to Nick Chater, now at, Department of Psychology, University of Edinburgh, 7, George Square, Edinburgh EH8 9JZ, Scotland, U.K.

A proof-theoretic account involves three components: the specification of (i) a formal language; (ii) a set of syntactic (*i.e.* proof-theoretic) rules of inference; and (iii) a mechanistic implementation of (i) and (ii). That is, cognition is an implemented formal logic. This is the classical, *logicist* position in cognitive science and artificial intelligence (Fodor & Pylyshyn, 1988; Hayes, 1978, 1984a). *Defeasible* inferences are inferences which can be *defeated* by additional information. Inferences licensed by classical logic are *monotonic*: no additional premises can invalidate a previously derived conclusion. This contrasts with everyday, defeasible inference which is *non-monotonic*: the addition of premises may invalidate a previously derived conclusion. In defeasible, non-monotonic inference it is possible to *add* premises and *lose* conclusions. Defeasible inference permeates every area of cognitive activity. Thus, at least *prima facie*, a logicist account of cognition must postulate proof-theoretic rules defined for some *non-monotonic* logic. We assess the practical attempt, in AI knowledge representation, to carry out this logicist programme using non-monotonic logics. We note that such logics are able to draw only unacceptably weak disjunctive conclusions; and that the theorem-proving algorithms over such logics are computationally intractable due to their reliance on solving the NP-complete problem of consistency checking. We suggest that the programme of logicist cognitive science is infeasible, and reply to a number of plausible objections to this conclusion.

The structure of the paper is as follows. We first characterise the classical, logicist position, using the formulation of two of its most influential exponents, Jerry Fodor and Zenon Pylyshyn, and adduce various adequacy criteria on logicist explanations of cognitive phenomena. We then note that human inferential processes, in common-sense reasoning, and in a variety of specific cognitive domains, are quite generally knowledge-rich and defeasible. These difficulties infect logicist treatments invoking unconscious, implicit inferences in text comprehension, conceptual reasoning, problem solving, perception, and even in recent accounts of human performance on explicit deductive reasoning tasks. To further illustrate the nature of the problem, we then draw on a parallel between these difficulties and those experienced in the philosophy of science in attempting to provide a *theory of confirmation*, a parallel also noted by Fodor (1983) (see also Sperber & Wilson, 1986). A specific attempt to deal with defeasible inference using non-monotonic logics (Reiter, 1985), which has been proposed within the tradition of knowledge representation in AI, is then critically examined. We draw the general moral that non-monotonic logics licence only unacceptably weak conclusions, and cannot be computationally implemented in real time. There are a number of proposals which appear to circumvent these problems. However, we argue, case by case, that such proposed logicist solutions succumb to the difficulties that we raise or amount to a retreat from the logicist position, and conclude that logicist cognitive science is ill-founded.

2. Logicist Cognitive Science

Fodor and Pylyshyn (Pylyshyn, 1973, 1984; Fodor, 1975, 1980, 1983, 1987; Fodor, Bever & Garrett, 1974; Fodor & Pylyshyn, 1988) have, over a number of years, argued that folk-psychological explanation, in terms of the ascription of propositional attitudes such as beliefs and desires, must be reconstructed in any proper account of cognitive activity. According to this view, to have a propositional attitude is to stand in a certain relation (the relation of believing, desiring or whatever) to a mental representation, the content of which is the object of the propositional attitude. Since the contents of propositional attitudes are described in natural language, the interpretation of the corresponding mental representations must be at the level of everyday objects and relations. This is the substance of the Representational Theory of Mind (*e.g.* Fodor, 1980). Folk psychology explains behaviour in terms of inference over propositional attitudes. Hence, a representationalist reconstruction of folk psychology must provide mechanisms for drawing inferences over the representations which capture the content of the propositional attitudes. These mechanisms are typically taken to be formal operations over syntactically structured representations. That is, mental operations are taken to apply purely in virtue of the structural properties of the representations. These syntactic mental operations must be coherent with respect to the semantics of the representations being manipulated. This is the substance of the Computational Theory of Mind (Fodor, 1980). Currently, the only way in which the semantic coherence of formal structural manipulation may be guaranteed is by showing that each manipulation of the representations corresponds to a sound proof-theoretic derivation in some appropriately interpreted formal language. In other words, the language of mental representation constitutes a *logic*, in which mental representations correspond to well-formed-formulae, and manipulations over them correspond to sound logical inferences. According to this view, a central task of cognitive science is to characterise the logical language of mental representation, the proof theoretic rules defined over it, and the content of the representations employed in the production of particular behaviours.

For the logicist, this provides a complete *psychological* explanation of performance of those behaviours. This proof-theoretic psychological explanation is *autonomous* (Fodor & Pylyshyn, 1988, p. 66; Chater & Oaksford, 1990) from the biological substrate underlying perception, memory, action and so on. Lower level biological explanations are taken to be independent from, and to fall outside the domain of, *psychological* explanation. This position may be elucidated by considering the three levels of explanation that David Marr (1982) took to constitute a complete account of the performance of a cognitive task. The claim that cognition is proof theory amounts to a restriction on the form of the level 1 (computational) theory. That is, a computational theory of some task must be specified (or at least must

be *specifiable*) as a proof theory over some interpreted logical language, and particular representations used in the performance of the task. Further, the logicist position also places restrictions on the form of the level 2 (algorithmic) theory. That is, it must characterise the theorem proving mechanism which animates the proof theory. This *theorem-prover* instantiates the control regime which determines which inferences are made when, in the performance of the task. This mechanism is defined over the formal properties of the logical expressions over which it is operating. It is these first two levels which the logicist takes to constitute *psychological* explanation. A level 3 (implementational) theory should constitute an account of how the theorem-prover specified at level 2 is instantiated in biological hardware. For the logicist this level is below, and largely independent of, the level of psychological explanation.

The classical, logicist cognitive science picture may be decomposed into four claims:

1. Cognition is computation.
2. Computation is formal.
3. Formal computation is mechanised proof theory.
4. The internal language over which the proof theory is defined is interpretable at the level of everyday objects and relations.

Within the framework of cognitive science, claim 1 must surely be taken as axiomatic. There is, however, substantial room for debate about the implications and status of claims 2, 3 and 4.

2. Computation is formal. That is, computational processes operate purely in terms of the *form*, syntax or shape of the symbolic structures over which they are defined. For example, consider the formal inferences of modus ponens (MP) and modus tollens (MT):

$$\text{MP: } p \rightarrow q,\ p \vdash q$$

$$\text{MT: } p \rightarrow q,\ \neg q \vdash \neg p.$$

These may be computed without reference to the meanings of the propositions p, q or the meaning of the connective $\rightarrow$. The premise '$p \rightarrow q$' is not treated as an atomic, unstructured lump, but as having syntactic structure: as having 'p', '$\rightarrow$', and 'q' as constituents. From the point of view of formal computation, all that matters for the application of *modus ponens* is that the 'p' in the second premise has the same shape as the 'p' on the left hand side of the first premise; and that the 'q' of the conclusion has the same shape as the 'q' on the right hand side of the first premise. This applies, *mutatis mutandis*, for *modus tollens*. Formal processes need not, as in this case, involve logical inference. List manipulation, sorting algorithms, sequences of procedural instructions etc. all count as formal, since they are defined over the shape rather than the content of their inputs.

Given the wide range of processes and schemes which have been taken to be computational models of cognition, the claim that computation is formal is not strictly true. Or rather, the requirement that computation is formal is *prescriptive* of the way in which Fodor and Pylyshyn (1988) would like the term 'computation' to be used, rather than *descriptive* of the way in which it *is* used in the range of literatures involved with mechanistic models of thought. For example, the mechanism of holographic memory, analogue computational methods, genetic learning algorithms and connectionism are not syntactic—the representations over which they operate typically *have* no syntactic structure. A possible confusion may arise, since these computational mechanisms can be simulated to an arbitrary degree of accuracy (and in some cases, perfectly) by the formal operations of a digital computer. However, any system (formal or otherwise) can be represented by formulae in some formal language, and its behaviour modelled by structure-sensitive operations over those formulae. It is in virtue of this fact that the general purpose digital computer is *general purpose*.

3. Formal computation is mechanised proof theory. Relevant information is represented as a set of formulae in a logical language, and computation proceeds by the operation of a theorem-prover for that language. The theorem-prover decides which proof-theoretic rule to apply when. *Prima facie*, theorem proving is a very particular form of computation. Again a possible confusion arises, since any computation can be *simulated* by the operation of an appropriate theorem-prover. Since any computation can be simulated on a Turing machine, for any computer program, there will be a corresponding Turing machine, with identical input–output behaviour. Any Turing machine can be axiomatised in first order logic (Boolos & Jeffrey, 1980), and hence any computation can be implemented on a theorem-prover for first order logic. Although any computation *can* be implemented in this way, almost invariably they are not.

4. The internal language over which the proof theory is defined is interpretable at the level of everyday objects and relations. The formulae over which the proof theory operates could, in principle, have an arbitrary semantics. Since the logicist takes propositional attitudes to be relations to these formulae, the contents of (at least some of) these formulae must correspond to the objects of beliefs and desires. In particular, therefore, the semantics of these formulae will make reference to everyday objects and properties—to tables, chairs, people, colours, feelings, and so on. It is hence unsurprising that the atomic terms of knowledge representation formalisms in artificial intelligence and cognitive psychology (such as semantic nets, schemas, production rules, and so on) stand in close correspondence with the lexical items of natural language. Indeed, for the sake of transparency, the atomic terms in AI knowledge representation are typically borrowed from the vocabulary of natural language. For example,

a program which encodes knowledge about an average taxpayer might start as follows (Clocksin & Mellish, 1984, p. 87):

```
average_taxpayer (X) :-
        not(foreigner(X)),
        not((spouse(X,Y), gross_income(Y,Inc),Inc>3000)),
        gross_income(X,Inc), ...
```

Of course, the logicist is not restricted to postulating representations defined at the level of tables and chairs. For the purposes of modelling specific cognitive processes, such as language understanding, perception and so on, the interpretations of the symbols may be phonetic, phonemic or syntactic categories, auditory and visual features, and the like.

The conjunction of assumptions 2, 3 and 4 constitutes a strong hypothesis about the nature of mental representations and mental processes. Having characterised the logicist picture, we now discuss certain adequacy criteria to which such an account should, at least in principle, be able to conform.

3. *Adequacy Criteria for Logicist Explanation*

We shall outline two main adequacy criteria which the logicist programme must be able to meet. Firstly, the proof-theoretic rules of inference defined over the postulated logical language (or languages) are capable of characterising the inferences implicated in human cognition. That is, the proof-theoretic rules must capture what we take pre-theoretically to be the semantically appropriate defeasible inferences. In Susan Haack's (1978) terminology, the logic(s) should be capable of respecting the appropriate *depraved semantics* (Haack, 1978, p. 188). So in the case of a non-monotonic logic for defeasible reasoning, the interpretation of the formalism must map appropriately onto our common sense or *depraved* understanding of defeasible inference. Some suitable non-monotonic logic must therefore capture the range of inferences which common sense licenses or, in other words, it should be complete with respect to the depraved semantics. By loose analogy with the notion of completeness in classical logic with respect to a standard formal semantic interpretation, we shall call this the *completeness** criterion. So a *complete** logicist explanation in some domain must provide a logical language and set of inferential rules which at least roughly captures our intuitions about inference in that domain.

Secondly, logicist explanation should, in principle, be able to provide a unified account of the cognitive processes within some domain, which covers each of Marr's (1982) explanatory levels. We shall call the constraint that such a unified explanation can be provided the *coherence* criterion. A *coherent* logicist account would provide a specification of a logical language in which knowledge is represented, and a proof theory defined over that language (level 1); a theorem-prover for that proof theory (level 2); and an

explication of how that theorem-prover is implemented in the brain (level 3).

A coherent logicist explanation must, among other things, be able to provide a level 2 algorithm appropriate to the level 1 proof theory, and to implement the level 2 algorithm in neural hardware. In practice the logicist is wont to insist that these relationships need not constrain theorising at each of the 3 levels. Indeed, one of the methodological appeals of the logicist view is that the implicit independence of each of the levels appears to licence the pursuit of high-level cognitive theorising, while we remain in comparative ignorance of the operation of the brain. This tenet of the logicist view (Fodor & Pylyshyn, 1988) presumably depends upon the following reasoning. Neural hardware (level 3) is surely able to implement such a simple symbolic device as a Turing machine or equivalent. But since any computable algorithm can be computed by a Turing machine, neural hardware appears to place no constraint at all on the algorithms which are psychologically plausible. Moreover, as long as the level 1 theory of the task domain, specified in terms of a set of proof theoretic axioms, is decidable, then there will be many level 2 algorithms for performing the task. By the previous argument, any such algorithm must be implementable at level 3, since any algorithm can be implemented in a Turing machine. So, according to the logicist, psychological explanations at levels 1 and 2 are relatively independent, that is they are *autonomous* (Chater & Oaksford, 1990) from the (level 3) biological substrate.

This line of reasoning may be taken to establish that almost any explanations postulated at each of the three levels are likely to be *in principle* compatible. To establish that logicist explanation is *coherent*, this weak, in principle, compatibility between explanatory levels must be supplemented by a strong *in practice* compatibility. That is, the level 1 proof theory must have a level 2 theorem proving algorithm which is not just computable but *computationally tractable*. Moreover, this level 2 algorithm must be able to run (level 3) on biological hardware with real-time characteristics compatible with the speed and effectiveness of observed behaviour. Indeed, only given a unified explanation of each of these levels can precise psychological predictions be made about the character of real-time performance.

The mere fact that we have a decidable set of proof-theoretic axioms (at level 1), guarantees only that there is a computable theorem proving algorithm; it does not guarantee that any such algorithm is computationally tractable. *In principle* computability results are sadly no guide to practical computational feasibility. Moreover, although any computable algorithm can be implemented on a Turing machine, and although the biological substrate is able to implement an arbitrary Turing machine, the nature of the biological substrate and the way in which the algorithm is implemented in that substrate will crucially affect the run time of the algorithm. Hence the nature of the hardware of the brain may considerably constrain the class of psychologically plausible algorithms.

Hence there are two species of doubt which may be raised concerning the coherence of the logicist programme. Firstly, it may be doubted that it is possible to implement theorem-proving algorithms postulated by the logicist in biological hardware such that they satisfy the real-time processing characteristics of cognitive performance. Secondly, in many psychological tasks, it may be doubted that there exists a tractable level 2 theorem-proving algorithm which instantiates the postulated level 1 theory. We have argued elsewhere (Chater & Oaksford, 1990) that the first species of doubt, the constraint that level 2 algorithms must be biologically implemented, militates strongly against the feasibility of an autonomous Logicist account. In this paper, with regard to the *coherence* of the logicist position, we concentrate on the second of these concerns: tractability. We argue that there may be no tractable algorithms appropriate to the level 1 theory which the logicist is forced to postulate. Moreover, logicist explanation must be not only tractable but *complete**. That is, the level 1 theory must actually be able to account for human inferential processes. We shall argue that the logicist account is also inadequate in this regard: it seems unlikely that a proof-theoretic level 1 account of human inferential processes will be forthcoming.

Since we are arguing against logicist approaches to cognition on the grounds that they may be unable to account for the defeasibility of human inference, it is incumbent upon us to show that human inference *is* defeasible, across a range of cognitive domains. It is to this task that we now turn.

4. The Defeasibility of Human Inference

Human knowledge is inherently revisable—expectations are routinely disconfirmed, norms violated, and what is certain today is discredited tomorrow. Human knowledge is also invariably partial and inferences must be drawn on the fly with incomplete knowledge of the relevant facts. The ability to reason and act appropriately in the face of overwhelming ignorance is one of human cognition's most remarkable and important achievements, and poses one of psychology's greatest challenges. In a mysterious and changing world, every conclusion is revisable and every premise open to question.

Consider, for example, the process of boiling an egg. Perhaps Egon has learnt from experience that if he puts an egg in boiling water then five minutes later the egg will be medium boiled. Having put the egg in the water as usual, Egon infers that the egg will be ready for his breakfast in five minutes. Such inferences, however, are radically defeasible. After all, there might be a power failure, an earthquake, Egon's careless brother may upset the pan, there may be salt in the water, the egg may be at altitude in an Everest base-camp, and so on. In these situations, Egon's inference that the egg will be ready to eat in five minutes time will be defeated.

Such inference is difficult to capture within a proof-theoretic framework. It is a feature of most standard logics that if a conclusion follows from some set of premises, then it still follows when additional premises are added. Logics in which this property holds are *monotonic* logics. Such logics are, at least *prima facie*, inappropriate for modelling inference in examples such as the above. According to a monotonic system of inference, if Egon infers that his egg will be ready five minutes after putting it in the boiling water he will be unable to revise this conclusion. So, for example, he must necessarily continue to expect his egg to be medium boiled even after his brother has knocked over the pan. In other words, if Egon were to reason according to a monotonic logic, then he would be unable to revise his tentative conclusions however strong the evidence to the contrary. This appears to imply that the proof theory that the logicist must postulate to deal with common-sense reasoning must be *non-monotonic*.

Non-monotonicity is required to model not just examples such as the above, but to capture non-demonstrative inference in general. Consider, for example, inductive reasoning, in which a general rule must be derived from a set of specific instances. This mode of reasoning is notoriously non-monotonic—however many premises of the form 'Raven A is black', 'Raven B is black' *etc.* are entertained, the inductive conclusion that 'All ravens are black' may be defeated by a single additional premise 'Raven N is white'. The defeasibility of induction has led many to doubt that induction is a justifiable species of inference at all. Whether or not induction is philosophically justifiable, people manifestly induce general laws on which to base their reasoning and action, from specific observations. So, whether or not there is a *philosophical* theory of induction, there must be a *psychological* theory of induction. Moreover, for classical, logicist cognitive science the form of this theory must be proof-theoretic. That is, for the logicist, induction, and all other species of non-demonstrative inference, must be assimilated to deduction.

In philosophy, other forms of non-demonstrative inference are typically seen as derivative on induction (Peirce, 1931–1958). In the above example, we assumed that Egon had induced the law that putting the egg in boiling water results in a medium boiled egg five minutes later. Having put a particular egg in boiling water, he applies this law to make the specific prediction that the egg will be medium boiled in five minutes. An inference from a particular occurrence of the antecedent of an inductive law, to a particular occurrence of the consequent of that law, is known as *eductive* inference. As we noted above, eductive inference, like inductive inference, is non-monotonic. Similarly, Egon's brother, who has also induced this law, may infer that the egg was put in boiling water five minutes earlier, from the fact that Egon is about to eat a medium boiled egg. Such an inference from a particular occurrence of the *consequent* of an inductive

law, to a particular occurrence of the *antecedent* of that law, is known as *abductive* inference or inference to the best explanation. Abductive inference is again notoriously non-monotonic. That the egg is medium boiled does not necessarily mean that it must have been in boiling water for five minutes—Egon may have boiled it for two minutes in the pressure cooker.

These non-monotonic modes of inference are implicated throughout almost every area of cognitive activity. The implicit inferences underlying text comprehension depend on the application of prior world knowledge to fill out and elaborate the information given in the text (Bransford & Johnson, 1972, 1973; Bransford, Barclay & Franks, 1972; Bransford & McCarell, 1975; Clark, 1977; Minsky, 1975; Stenning & Oaksford, 1989). All such implicit inferences can be defeated by subsequent sentences contradicting our implicit conclusions. Theories of concepts which are concerned to capture the family resemblance or prototype structure of human categorisation implicitly recognise the defeasibility of semantic knowledge. So, although not all birds can fly, the prototypical bird is represented as flying, the majority of exemplar birds fly, the probability that a bird flies is high, *etc.* depending on the theory that one considers (Rosch, 1973, 1975; Medin & Schaffer, 1978; Nosofsky, 1986). According to modern constructivist theories of perception, much of perceptual processing is taken to involve inference to the best explanation about the state of the environment, given perceptual evidence. The defeasibility of such inference is evidenced by the possibility of perceptual illusion and error (Gregory, 1977; Fodor & Pylyshyn, 1981; MacArthur, 1982). Non-demonstrative modes of inference have even been argued to encroach upon apparently deductive tasks such as conditional reasoning (Oaksford, 1988; Byrne, 1989; Oaksford, Chater & Stenning, 1990). Thus, the whole of cognitive performance depends upon non-monotonic inferential processes. If these cannot be elucidated within the logicist, proof-theoretic framework, then almost every interesting cognitive phenomenon will fall outside the scope of logicist psychological explanation.

5. *Non-monotonicity and Confirmation in Science*

Prima facie, the logicist programme is analogous to the Logical Positivist's attempts to provide a theory of confirmation for scientific theories (Carnap, 1923, 1950; Hempel, 1952, 1965). Roughly, it was hoped that such a theory could be axiomatised as an inductive logic, which has the form of deduction in reverse. The claim was that in induction a statement is confirmed by the truth of its deductive consequences, whereas in deduction the truth of a statement guarantees the truth of its deductive consequences. Unfortunately, the axioms of such putative inductive logics could not be made mutually consistent and generated many paradoxes. For example, from very minimal assumptions about the form of an inductive logic it is possible to prove that any hypothesis confirms any other

hypothesis (Goodman, 1983, originally 1954). The proof is trivial, and exploits the fact that confirmation appears to flow in both directions between hypotheses and their consequences. Consider two arbitrary hypotheses H and H′. The conjunction H ∧ H′ has H as a consequence, and hence, since confirmation is supposed to be deduction in reverse, H confirms H ∧ H′. If H ∧ H′ is true then H′ must be true—so according to any sensible confirmation theory surely H ∧ H′ must confirm H′. Indeed presumably the strength of this confirmation should be the greatest possible, since if H ∧ H′ is true, then H′ is definitely true—*i.e.* maximally confirmed. We have concluded that H confirms H ∧ H′ and H ∧ H′ confirms H′. Assuming transitivity, which again seems necessary for any inductive logic able to support the elaborate chains of confirmation in science, this means that H confirms H′ (and, of course *vice versa*). Since H and H′ were chosen arbitrarily, we have the paradoxical conclusion that any two hypotheses confirm each other.

Further, Goodman's (1983) famous 'grue' predicate

$$(\forall x(x \text{ is grue at } t \longleftrightarrow (x \text{ is green } \& \ t < \text{year 2000}) \vee (x \text{ is blue } \& \ t \geq \text{year 2000})))$$

showed that the problems of confirmation theory could not be resolved by purely formal considerations. Every emerald which has so far been observed is both grue and green. Yet the induction to *all emeralds are green* will continue to be true after the year 2000, whereas the induction to *all emeralds are grue* will clearly fail from the year 2000, after which no emeralds will be grue. In Goodman's terms, 'green' is a projectible predicate where 'grue' is not. The projectibility of predicates such as 'green' and the non-projectibility of predicates such as 'grue' could not inhere in their formal properties; the projectibility of a property could not be dependent on the shape of the predicate symbol used to denote it!

Fodor (1983) raises further problems for the procedures of inductive confirmation: such non-demonstrative fixation of belief is both *isotropic* and *Quinean*.

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

> By saying that scientific confirmation is Quinean, I mean that the degree of confirmation assigned to any given hypothesis is sensitive to properties of the entire belief system. (p. 107)

That confirmation is Quinean is indicated by criteria of theory preference

which are based on global properties of a system of scientific beliefs. Properties such as simplicity, plausibility, conservatism or projectibility (see above) are global properties in just this sense. Fodor (1983) argues that such global properties cannot be handled by any current theory of confirmation—and that, in consequence, there is no serious theory of scientific confirmation.

The failure of a logicist account of science does not, of course, necessarily entail that a logicist account of mind will be similarly unsuccessful. However, there is reason to suppose that ordinary everyday common-sense inference may be relevantly analogous to confirmation in science, and hence that a logicist account of one may stand or fall with a logicist account of the other. Jerry Fodor (1983), although a staunch advocate of a proof-theoretic account of mind, argues for the analogy very eloquently. He notes that the problem of confirmation in science maps rather directly onto the everyday, commonsense reasoning problem of knowing how to update one's beliefs, given that one has performed some action—the notorious, and ubiquitous *frame problem* in AI. Fodor considers the predicament of an artificial robot acting on the world, and trying to revise its beliefs appropriately in consequence:

> How . . . does the machine's program determine which beliefs the robot ought to reevaluate given that it has embarked upon some or other course of action? What makes the problem so hard is precisely that it seems unlikely that any *local* solution will do. . . . the following truths seem to be self-evident: First, that there is no fixed set of beliefs . . . that . . . are the [only] ones that require reconsideration . . . Second, new beliefs don't come docketed with information about which old beliefs they ought to affect . . . Third, the set of beliefs apt for reconsideration cannot be determined by reference to the recency of their acquisition, or by reference to their generality, or by reference to merely semantic relations between the contents of the belief and the description under which the action is performed . . . etc. Should any of these propositions seem *less* than self-evident, consider the special case of the frame problem where the robot is a mechanical scientist and the action performed is an experiment. Here the question 'which of my beliefs ought I to reconsider given the possible consequences of my action' is transparently equivalent to the question "What, in general, is the optimal adjustment of my beliefs to my experiences?". This is, of course, exactly the question that a theory of confirmation is supposed to answer. (Fodor, 1983, p. 114)

The frame problem is simply a particular example of a problem in which defeasible, non-demonstrative inference must be performed in a knowledge-rich domain.

> . . . as soon as we begin to look at . . . processes . . . of non-demonstrative fixation of belief we run into problems that have a quite characteristic property. They seem to involve isotropic and Quinean computations; computations that are . . . sensitive to the whole belief system. This is exactly what one would expect on the assumption that non-demonstrative fixation of belief really is quite like scientific confirmation, and that scientific confirmation is itself characteristically Quinean and isotropic. (Fodor, 1983, pp. 114–5)

Of course, Fodor, couches his discussion in terms of the fixation of *belief*. The same difficulties will arise for the management of any data-base over a knowledge rich domain, whether or not the statements in that data-base may appropriately be interpreted as beliefs.[1]

Let us sum up the argument so far. Quite generally, it seems that in domains in which mental processes are held to be inferential, that inference will typically be non-demonstrative, defeasible inference. Hence the challenge of modelling non-demonstrative inference within a proof-theoretic framework is central to the feasibility of a logicist account. Yet the failure of Logical Positivism to assimilate non-demonstrative inference to a deductive framework, the failure to devise a successful inductive logic, the inability to account logically for scientific knowledge and theory change, and the like, raise the suspicion that the logicist programme in cognitive science and artificial intelligence may be unworkable. The analogy with the philosophy of science serves to indicate the magnitude of the problem confronting researchers who are attempting to develop non-monotonic logics.

Suggestive as such general theoretical considerations are, the proof of the logicist pudding is, of course, entirely in the eating. If the logicist framework does appear to provide a plausible account of defeasible inferential processes, then the general theoretical qualms that we have raised

---

[1] Fodor is concerned to outline an interesting and important distinction—between *central* processes of non-demonstrative belief fixation, which are Quinean and isotropic; and domain specific processes, in which the inferential processes are not dependent on the whole belief system, but only on a prescribed set of information, relevant to that domain. Fodor takes the demarcation between the former *central* processes and the latter *informationally encapsulated* processes to distinguish areas in which cognitive science is likely to prove infeasible from areas in which progress may be made. Note that domain specific systems may involve non-demonstrative inference, and that this inference may be Quinean and isotropic relative to all the knowledge encoded in the module. So the nondemonstrative defeasible inference that appears to be implicated in putatively domain specific processes involved in language understanding and perception may be just as problematic as the central processes of common-sense inference. With regard to our concern in this paper, the key distinction is not between domain-specific and central processes but between processes which involve knowledge-rich defeasible inference and are at least *prima facie* problematic for a logicist account, and those which do not. Of course, it is possible that this distinction is in practice rather trivial, all human inference being of the former kind.

may be put aside. Moreover, profound and heretofore unrealised implications for the philosophy of science would result. In the following section we therefore examine the current stage of the logicist attempt to account for defeasibility, as embodied in the field of knowledge representation in AI, and argue (i) that logicist accounts fail, and (ii) that they fail in principled ways.

Firstly, the proof-theoretic rules for the non-monotonic logics that have been proposed to capture defeasibility do not adequately capture knowledge-rich human non-demonstrative inference—using the terminology that we introduced above, such logics are not *complete**. In particular, non-monotonic logic appears able to generate only unacceptably weak disjunctive conclusions. Secondly, such non-monotonic logics do not possess any tractable algorithms—that is, the computational resources required by theorem-provers for such logics increase explosively as the number of formulae over which we must reason increases. *Prima facie*, this appears to rule out a proof-theoretic view of cognition for domains in which a large amount of knowledge must be taken into account. In short, the proof-theoretic account of defeasibility does not give the right inferential behaviour, and is computationally intractable. Given the extent to which almost every cognitive task involves defeasible, non-demonstrative inference, the domain of the proof-theoretic account may perhaps be unexpectedly limited.

6. *Artificial Intelligence and the Logicist Approach to Defeasible Inference*

A central challenge of logicist cognitive science is to provide a proof theory and theorem-proving methods which capture non-monotonic inference. Workers in artificial intelligence have faced this challenge most directly, in attempting to build systems which can reason about real-world, common-sense domains, using mechanised proof theory (for a general introduction to this approach, see Charniak & McDermott, 1985). In this section, we discuss two difficulties with this approach. Firstly, that non-monotonic inference licences only unacceptably weak conclusions; and secondly, that such theorem proving for such logics is computationally intractable.

In order to cope with the defeasibility of inferential rules in examples such as the above, it is necessary to devise a logical scheme in which defeasible rules may be encoded. A wide variety of superficially very different non-monotonic logics have been proposed. The best known are McCarthy's (1980) *circumscription*, Reiter's *default logic* (1980, 1985), McDermott and Doyle's (1980) *non-monotonic logic I*, McDermott's *non-monotonic logic II* (1982), and Clark's *predicate completion* (1978). The problems that we shall raise appear to apply equally to all of these approaches (Hanks & McDermott, 1985, 1986; Shoam, 1987, 1988).

6.1 Non-monotonic Logics and Weak Conclusions

For concreteness we shall consider a formalisation of defeasible inference which introduces a meta-theoretic M operator into the object language of a standard logic (Reiter, 1980, 1985). Defeasible rules (in AI terminology, default rules) are encoded as follows:

$$\phi \wedge M\varphi \rightarrow \psi$$

This formula reads: ψ can be inferred from ϕ as long as $\neg\psi$ is not provable, given the axioms of the system. So the intuitive interpretation of $M\varphi$ is that $\neg\psi$ cannot be proved given Γ (the set of logical axioms which govern the behaviour of the connectives) and Δ (the non-logical axioms which encode the domain specific knowledge of the system). In other words, it is *consistent* to infer ψ from $\Gamma \cup \Delta$ and ϕ. The M operator has the unusual property of introducing the meta-theoretic concept of deducibility ($\vdash$) into the object language—*i.e.* $M\psi$ is equivalent to $\Gamma \cup \Delta \nvdash \neg\psi$. (This logically inelegant manoeuvre may be avoided by interpreting the M operator as a modal operator, and providing a possible worlds semantics for the resulting logic (McDermott & Doyle, 1980). Which formulation is used makes no difference to the inferences that can be drawn, or to the theorem-proving algorithms employed.)

Returning to our example of Egon and the egg, suppose that Egon tells his brother that he has just put an egg in boiling water. Egon's brother's relevant prior knowledge may be encoded in axioms (Δ) of something like the following form:

1. ($\mathbf{egg}_i$) in boiling water at t $\wedge$ M(($\mathbf{egg}_i$) medium boiled at t + 5 minutes) $\rightarrow$ ($\mathbf{egg}_i$) medium boiled at t + 5 minutes

2. ($\mathbf{egg}_i$) in pressure cooker at t $\wedge$ M((egg_i) hard boiled at t + 5 minutes) $\rightarrow$ ($\mathbf{egg}_i$) hard boiled at t + 5 minutes

3. ($\mathbf{egg}_i$) medium boiled at t + 5 $\rightarrow$ not ($\mathbf{egg}_i$) hard boiled at t + 5

4. ($\mathbf{egg}_i$) hard boiled at t + 5 $\rightarrow$ not ($\mathbf{egg}_i$) medium boiled at t + 5

(The non-default premises 3 and 4 simply encode the fact that an egg can not be both hard boiled and medium boiled at the same time)

He now knows that a particular egg (egg_1) is in boiling water and adds 5 to Δ:

5. egg_1 in boiling water

Since, 5 matches the first conjunct of the antecedent of 1 the possibility arises that egg_1 will be medium boiled at t + 5. Since, it is not possible to derive the negation of this proposition from 1 to 5, then this conclusion is consistent with the data base—M(egg_1 medium boiled at t + 5 minutes)—

the second conjunct of the antecedent of 1 is also satisfied. So, the consequence that this egg will be medium boiled in five minutes may legitimately be inferred.

Egon's brother now walks into the kitchen, and observes that the egg must be in the pressure cooker (it is the only pan on the stove). In our formalism, this amounts to adding 6 to 1–4.

6. egg_1 in pressure cooker

Since 6 matches the first conjunct of the antecedent of 2 the possibility arises that egg_1 will be hard boiled at t + 5. Since it is not possible to derive the negation of this proposition from 1–4 and 6, then this conclusion is consistent with the data base—M(egg_1 hard boiled at t + 5 minutes)—the second conjunct of the antecedent of 2 is also satisfied. So, the consequence that this egg will be hard boiled in five minutes may legitimately be inferred.

Yet this situation may seem paradoxical. From 1–4 and 5 we have the conclusion that the egg is medium boiled at t + 5 (and hence, by 3, it is not hard boiled). On the other hand, from 1–4 and 6 we have the conclusion that the egg is hard boiled at t + 5 (and hence, by 4, it is not medium boiled). This may seem counterintuitive if we are used to monotonic logics. For in such a logic all the conclusions that follow from any subset of 1–6 must follow from the complete set. In particular, 1–6 would imply that the egg is both hard boiled and not hard boiled—that is, the axioms are inconsistent. However, since the logic is non-monotonic, inconsistency does not follow.

The cases in which the egg is medium boiled and hard boiled are what are known as distinct *extensions* of 1–4. Which extension is obtained depends on which default rule is used first. If rule 1 is used first to infer that the egg is medium boiled, rule 3 can be used to infer that it is not hard boiled. In this extension, it is inconsistent to assume that the egg is hard boiled—that is M(($\mathbf{egg}_i$) hard boiled at t + 5 minutes) cannot be satisfied, the contrary default rule 2 is blocked, and hence no contradiction results. Similarly, we can consider the extension in which rule 2 is used first. In this case, the egg is inferred to be hard boiled, and hence, by rule 4, it cannot be medium boiled. Thus, rule 1 cannot apply, and no contradictory conclusion is derived. Given that there are two possible extensions of 1–6, what conclusions can be derived? The only valid conclusions are those that hold in *all* extensions—so rather than inferring any particular extension, we may infer only the *disjunction* of all extensions. In the present case, this is simply that:

7. egg_1 hard boiled V egg_1 medium boiled

This disjunctive conclusion is not intuitively adequate (that default logics give only such weak conclusions amounts to what McDermott (1986) calls

the 'you don't want to know' problem. From the point of view of prediction and action, *you don't want to know* that the egg will be either medium or hard boiled—you want to know which!). The performance of the system contrasts with human reasoning. If we know that the egg is in boiling water and that it is in the pressure cooker, then we will unambiguously infer that it will be hard boiled at t + 5. Whereas the system has no way of resolving conflicting default conclusions, at least in cases such as this, such resolution is an effortless feature of human cognition. Hence, to model human performance, the system must be able to determine how conflicting pieces of inconclusive evidence bear upon the inferences that may be drawn. In other words, the system must solve the problem of appropriately revising its beliefs in the face of incomplete and conflicting information. Yet this *is* the problem of non-demonstrative inference. So in trying to explain non-demonstrative inference, by invoking non-monotonic logics, we have succeeded only in raising it again. Given the failure of Logical Positivist attempts to reconstruct non-demonstrative inference proof-theoretically, perhaps the failure of AI to tackle the same problem is unsurprising.

Despite this worrying state of affairs, within the AI community there have been attempts to tackle the problem of resolving incomplete and conflicting evidence by using domain-specific heuristics. Such heuristics are intended to differentiate acceptable from unacceptable extensions of the logical system. In view of the generality of the problem which such heuristics are attempting to solve, it is not surprising that they have been criticised as inadequate (Hanks & McDermott, 1985; Israel, 1980). Moreover, insofar as cognitive processes are taken to be semantically justified—*i.e.* to correspond to valid derivations at the level of proof theory—the postulation of such heuristics *in the control strategy of the theorem-prover* constitutes a retreat from the logicist position. However, let us assume that the problem of resolving conflicting and incomplete information could be solved by some set of heuristics. Even given this (apparently counterfactual) assumption, the logicist proof-theoretic programme appears to be infeasible.

6.2 Non-monotonic Logics and Computational Complexity

To complete the programme of Logicist Cognitive Science, it must be possible to construct tractable algorithms which embody the non-monotonic proof theory. In particular, the introduction of the M operator, or equivalent, requires the ability to check whether or not some premise is consistent with the current contents of the data-base ($\Gamma \cup \Delta$). Thus, *any invocation of a default rule requires a complete consistency check over the whole data-base*. However, as we shall see consistency checking is computationally intractable.

Consistency checking constitutes a general class of problems in complexity theory called *satisfiability* problems. In this section, we note the

intractability of such problems, and the consequent implausibility of the proof-theoretic account of non-demonstrative inference.

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

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

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

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

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

Unfortunately for the proof-theoretic programme of logicist cognitive science, consistency checking, like all satisfiability problems, is NP-complete. Hence an instantiation of a non-monotonic logic, which invokes a consistency check over the whole data-base every time a default rule is used, appears to be a hopelessly unpromising account of real-time defeasible human inference which is invoked rapidly and effortlessly in almost every cognitive task.

6.3. Do We Need to Appeal to Non-monotonicity?

We have argued against the logicist approach to cognitive science by showing that human inference is defeasible, that proof theory must therefore be defined for a non-monotonic logic, and that theorem proving for such a logic is incomplete* and intractable. The opponent of the proof-theoretic programme may agree with these points but argue that the appeal to non-monotonicity is unnecessary to defeat the logicist programme. In particular, it may be argued that computational intractability bites equally for standard, monotonic logics. After all, in almost *any* logic the general problem of deciding whether a given finite set of premises logically implies a given conclusion is NP-complete (Cook, 1971), and, of course, checking the validity of arguments is equivalent to checking the consistency of sets of propositions. According to this line of thought, the considerations of defeasibility and non-monotonicity that we have stressed appear to be wholly beside the point. However, there is a crucial difference between

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

the monotonic and non-monotonic cases. In monotonic logic, if a set of premises is consistent any application of a rule of inference will maintain consistency. This contrasts with the non-monotonic case, where each time a rule is applied, a new consistency check must be performed. So, if consistency checking is a problem for monotonic logics, it is a far greater problem for non-monotonic logics. Hence models of thought based on proof theory are severely undermined by the defeasibility of human inference, and the consequent postulation that the logic of thought must be non-monotonic. For the logicist, proof theory is supposed to be the *basis* of all cognitive activity (in common-sense reasoning, language, perception). If the logic of that proof theory is non-monotonic, and hence rule application is intractable, then the logicist position is surely untenable.

However, there are a number of possible logicist responses to this negative conclusion. We now consider these one by one.

7. Objections and Replies

7.1 Worst Case versus Typical Case

The *a priori* intractability results that we have considered are worst-case analyses. In practice the possibility remains that in typical cases, non-monotonic reasoning may be effected without exhausting the available computational resources. The most direct way to test this hypothesis is to perform an *a posteriori* analysis of actual *average-case* run-times of *implemented* non-monotonic logics. However, to the best of our knowledge, no such implementations exist. Of course, in computer science, theory is often developed in advance of its implementation in real systems. Such a situation is healthy if there is some reason to believe that implementations may be forthcoming—this does not appear to be the case in current approaches to defeasibility in the knowledge representation literature. This is of particular concern for artificial intelligence and cognitive science in which successful implementation is taken as the benchmark of theoretical rigour and adequacy. It is not, of course, possible to distinguish reliably between progressive and degenerating research programmes, between temporary puzzles for, and outright falsifications of, some line of research (Lakatos, 1970). However, increasing theoretical elaboration and decreasing practical success is surely a straw in the wind.

7.2 Heuristics, Tractability and Completeness*

Apart from the above, there is another reason why *a priori* intractability results are not necessarily taken to rule out the possibility of practical computation. No algorithm—*i.e.* no procedure that is *guaranteed* to solve the computational problem—may be tractable, and yet there may be more or less reliable *heuristics* which often solve the problem, or at least provide

something close enough to the solution to be useful. These heuristics need not necessarily be computationally intractable. Computational tractability may be bought at the price of the reliability of the procedures. Given that human inference is manifestly unreliable—we are always jumping to conclusions, forgetting to take into account important considerations, and so on—it may seem plausible that an appropriate set of heuristics may be the basis of human defeasible inference. In discussing heuristics as a method of solving a particular case of the problem of defeasible inference, the frame problem, Fodor says:

> The idea is that, while non-demonstrative confirmation (and hence, presumably the psychology of belief fixation) is isotropic and Quinean *in principle*, still, given a particular hypothesis, there are, in practice heuristic procedures for determining the range of effects its acceptance can have on the rest of one's beliefs. (Fodor, 1983, p. 115)

We noted above that such heuristics have been appealed to in the attempt to overcome the tendency of non-monotonic logics to give unavoidably weak disjunctive conclusions. Appropriate heuristics might, perhaps, systematically favour some possible extensions of knowledge-base over others—heuristics which take account of the structure of the world could, it may be hoped, show systematic bias in favour of what we intuitively consider to be the *right* extensions. Thus, the operation of the heuristics implicitly encodes knowledge about the world. This approach has indeed been pursued in the knowledge representation literature. Let us consider a famous problem in non-monotonic reasoning, the Yale shooting problem (Hanks & McDermott, 1985), and consider a heuristic designed to favour the 'right' answers.

A gun is loaded at some time, and fired at a person at some later time. The problem is to determine whether or not the person ceases to be alive. It is assumed that the firing of a loaded gun at a person is invariably fatal. Further, we assume two defeasible rules: that (i) if a gun is loaded at some time, then it will typically continue to be loaded at some later time; and (ii) if a person is alive at some time, that person will typically be alive at some later time. This scenario creates a problem analogous to the one we raised earlier with respect to Egon and the egg. For any non-monotonic or defeasible reasoning system two contrary, albeit defeasible, conclusions are warranted: either the person is not alive at some later time or he is alive at some later time (Hanks & McDermott, 1986). Observe that this example is a specific application of non-monotonic logics to the frame problem (see Fodor's comments quoted above). The scenario creates the problem of how to appropriately revise one's beliefs concerning the person being alive or dead given that a shooting has taken place.

Specific proposals concerning how to resolve the problem of multiple inconsistent extensions of a non-monotonic theory all invoke some method

of preferring one extension over another. Hanks and McDermott (1986) propose that if conclusions in two extensions are contraries, then an earlier defeasible conclusion should defeat later defeasible conclusions. Thus, in the Yale shooting problem, since rule (i) is invoked earlier than rule (ii) in the chain of reasoning, the intermediate defeasible conclusion that the gun is loaded when it is fired is to be preferred over the defeasible conclusion that the person is alive after the gun has been fired. This 'solution' is justified on the basis of reflections on the nature of causality (Shoam, 1986). However, although this move resolves the problem in favour of the putatively desired defeasible conclusion—that the person is dead at the later time—such a preference for one extension over another is not legitimised within the logical system. Moreover, Loui (1987) observes that although this heuristic may accord with intuition in the Yale shooting problem, there are many other examples where intuitions are violated if the heuristic is applied across the board. Thus, although such a temporal precedence heuristic may occasionally allow the right conclusion (although even this is disputed, see Loui, 1987), it is not guaranteed to do so.

Other methods for preferring one extension over another (*e.g.* Poole, 1985; Nute, 1985, 1986; Loui, 1986) all involve explicitly 'encoding the preference information' (Loui, 1987, p. 291). Thus the decision about what defeasible inferences are licensed is external to the inference regime, and reflect purely heuristic assumptions usually concerning the nature of causality. Relative to the isotropic nature of non-demonstrative inference it is doubtful whether any of these heuristic assumptions are of general applicability. Moreover, all of these assumptions are Quinean, they reflect global properties of our causal knowledge. However, in their implementation in non-monotonic logics they are imposed externally by the programmer. But to complete the proof-theoretic programme such properties need to be shown to emerge from the structure of our world knowledge and can not be imposed by fiat. Hence all these 'solutions' fail to be complete*.

It is important to note that the kind of heuristics proposed above to circumvent the incompleteness* of non-monotonic logics are distinct from the equally *non-logical* decisions enforced by any practical implementation of logic in for example PROLOG. Practical theorem proving requires various non-logical *control* decisions to be made in the search strategy of the theorem prover, for example, to employ backward chaining only, to use loop checkers and to employ the 'cut' operator (Hogger, 1984). These decisions involve the control strategy of an implementation of logic and as such are wholly independent of the knowledge to be encoded in a particular data-base. However, the heuristics proposed above specifically involve the *very knowledge* which is to be encoded. As we stated above, this involves making heuristic assumptions about how beliefs are appropriately updated. But this is precisely the problem which, on the proof-theoretic logicist account, non-monotonic logics were invoked to resolve! McDermott (1986) proposes a retreat to *proceduralism* in which it is admitted that no

semantic justification for the heuristics proposed will be forthcoming. We will discuss this option further below, but observe now that it directly contradicts Fodor and Pylyshyn's logicist account of cognitive science.

It seems that appeal to heuristics is unlikely to repair the incompleteness* of non-monotonic reasoning; and that, in any case, to the extent that world-knowledge is embodied in heuristics rather than represented in the logical language over which the proof theory is defined, the appeal to heuristics amounts to a rejection of the logicist account of inference. A further proposal, mentioned by Loui (1987), is to make reasoning *domain specific*. If only information *relevant* to a specific domain is employed in a particular inference then certain desirable consequences may follow. First, if a formal account of relevance can be defined, then it may be possible to logically delimit the sets of premises over which reasoning takes place. This *may* satisfy the completeness* criterion. Second, by restricting the premises to the relevant ones, n may be suitably restricted to satisfy the tractability criterion. We now turn to two proposals concerning the concept of *relevance*.

7.3 Relevance

Relevance logic restricts the concept of deducibility such as to avoid the well known paradoxes of material implication '$\supset$'. For example, it seems bizarre that $A \supset (B \supset A)$ is a theorem for arbitrary A and B, if '$\supset$' is held to capture an intuitive notion of implication. Anderson & Belnap (1975) define a notion of *relevant entailment* which employs a system of indices which attach to assumptions. The indices guarantee that a logical relation of relevance exists between the antecedent and consequent of a conditional statement. Only assumptions B, which rely on assumptions A, will allow '$\rightarrow$' (relevant entailment) to be introduced such that $A \rightarrow B$. That is, $A \rightarrow B$ can only be concluded when A is part of the subproof of B. In this precise logical sense A is relevant to B. It has been proposed by, for example Haack (1978), that this notion of relevant entailment could assist in avoiding the conclusion that confirmation is Quinean. Instead of the whole of scientific knowledge being the unit of confirmation she suggests that it could be just the *relevant* subset in Anderson and Belnap's sense. Moreover, Levesque (1988) proposes that relevant entailment may be used to effect a *tractable* selection of relevant premises from a data base for subsequent reasoning processes.

However, in reasoning in defeasible domains relevant entailment still violates both the completeness* and tractability criteria. Even supposing relevant entailment were employed, default rule application would still remain intractable (Levesque, 1988). There are also strong grounds to question whether relevant entailment is complete*. In introducing the complete* criterion we noted that formal concepts must respect the depraved semantics for the informal concepts they encode. However, it seems that the notion of deductive relevance captured in relevant entail-

ment far from exhausts the ways in which one piece of knowledge may be relevant to another piece of knowledge. First, Fodor (1983) observes that in science, knowledge in one domain may be relevant to another domain *analogically*. Strictly, considerations of analogical reasoning move outside the domain of confirmation into the domain of scientific discovery. For example,

> what's known about the flow of water gets borrowed to model the flow of electricity, what's known about the structure of the solar system gets borrowed to model the structure of the atom. (Fodor, 1983, p. 107)

However, analogical reasoning processes are part of our non-demonstrative reasoning abilities and as such require explanation by the mechanisms which purport to account for those abilities.

Second, relevant entailment accounts for relevance between propositions—it is a purely structural notion. However, our intuitions about relevance appear to be crucially dependent on lexical rather than structural properties of statements. For example, the fact that Fred having a heart is relevant to Fred's having palpitations depends not on the structure of the two propositions, but on the meaning of 'heart', 'palpitation' and the causal structure of the world which putatively links the two. Further, it appears that relevance is not determined by the *extension* of the relevant properties. According to the well-worn philosophical example, having a heart and having kidneys are co-extensive—so if Fred has either property he has them both. However, although Fred's having a heart may be relevant to his having palpitations, his having kidneys may not be.

This clearly suggests that 'relevance' is an *intensional* concept and hence it might be expected that a well-defined concept of relevance would be forthcoming via an appropriate possible worlds semantics. However, the provision of a proper semantics for relevance logics is notoriously difficult:

> The relevance logicians run the risk of turning logical validity into a clumsy thing. The difficulties they have in providing their largely proof-theoretic theories with a proper semantics may be regarded as a symptom of this. The semantic theories which have thus far been put forward tend to lack the explanatory power which is to be expected from theories which purport to say what relevance means. (Veltman, 1985, pp. 42–3)

In sum, it would appear that relevance logic fails to meet both our criteria. Default rule application remains intractable and there are grounds for considerable doubt over whether relevant entailment is sufficient to capture the numerous ways in which one piece of knowledge may be relevant to another piece of knowledge.

However, relevance logic does not exhaust attempts to define a notion

of relevance which may be of more general applicability. *Relevance Theory* (Sperber & Wilson, 1986) is an attempt to account for how a person's beliefs may be appropriately updated which takes a less restricted view of relevance and also incorporates various processing requirements which bear on the issue of tractability. Sperber and Wilson (1986) first emphasize a disanalogy between their account of the spontaneous and almost instantaneous updating of beliefs which occurs in sentence comprehension and the reflective and time consuming updating of beliefs which occurs in scientific theorising. It is the former which they are concerned to explicate. They suggest that the inferential processes underlying sentence comprehension must exploit only the *accessible* information. Sperber and Wilson (1986) then outline what we will term a *hybrid* inferential regime consisting of a restricted deductive mechanism and a non-logical component which is responsible for updating the *confirmation* strengths which attach to propositions stored in memory. The restricted logical component, which contains no introduction rules, is motivated primarily by issues of tractability but also represents a substantive claim about the nature of peoples' inferential processes in language comprehension. Sperber and Wilson (1986) are careful to emphasize that they *do not* intend their notion of confirmation strength to be conflated with the assignment of subjective probabilities to propositions which are explicitly manipulated in judging the relative strengths of those propositions. 'Confirmation strength' is to be understood as a purely processing notion determined by a proposition's prior history of being accessed from memory.

The concept of relevance is defined relative to a context C.

> **Extent condition 1**: an assumption is relevant in C to the extent that its contextual effects in C are large.
> **Extent condition 2**: an assumption is relevant in C to the extent that the effort required to process it in C is small.

Contextual *effects* and processing *effort* are defined in terms of the hybrid inferential regime introduced above. There is a trade off between these two 'extent conditions' in determining the relevance of an assumption.

The notion of relevance thus defined may not be helpful given our present concerns, since it appears to beg the very question we were hoping the concept of relevance would answer. That is, how do we choose from all we know the relevant items to update in response to new information. The above definition is relativised to a context C, which is understood as the old information available from the immediately prior discourse and from memory for encyclopaedic or world knowledge. Sperber and Wilson (1986, pp. 132–7) argue convincingly that the whole of the latter may be included in C, although it is suggested that this would violate extent conditions 1 and 2 of the definition. However, since relevance is defined in terms of C, delimiting C's extent by appeal to relevance would be viciously circular. Thus to avoid the charge of circularity independent

grounds are required to delimit C. Sperber and Wilson (1986, p. 138) appeal to the fact that in cognitive psychology and cognitive science knowledge is generally agreed to be compartmentalised into 'schemata', 'frames', 'scenarios', and 'prototypes'. However, it was precisely in search of principled grounds for this compartmentalisation that we embarked upon this discussion of relevance! 'Schemata', 'frames', 'scenarios' and 'prototypes' are precisely the names appropriated to the domain specific units of knowledge which it was hoped that the concept of relevance would provide thereby delimiting the isotropy of confirmation. It seems, therefore, that relevance theory, in order to define a restricted notion of relevance appropriate to sentence comprehension, must presuppose a solution to the more global problem of relevance, which is our present concern.

Apart from this, there are general problems for relevance theory. We will mention just two. First, Sperber and Wilson's (1986) account of their inferential mechanism seems to leave no room for *errors* of interpretation. These must be possible since the assumptions recruited from encyclopaedic memory in discourse are often of a defeasible elaborative form (Stenning & Oaksford, 1989). Such elaborative inferences can be defeated by subsequent discourse, and hence must be cancelled. This of course suggests that the logic of the inferential component is going to be non-monotonic even in sentence comprehension. Thus although introduction rules have been excluded to the benefit of the system's tractability, default rules will have to be included which, as we have seen, are unlikely to enhance the tractability of the system. Second, how the confirmation strengths are used and updated is currently opaque. The proposal is that as a proposition in memory is accessed more often so its ability to be accessed is enhanced. Thus its strength does not have to be explicitly represented. However, in a symbolic, deductive system, on the lines Sperber and Wilson (1986) propose, we can see no way of implementing this proposal. In a symbol system it matters not one jot how often an item is accessed, every time it is accessed it will be accessed in the order dictated by the program—*unless* some parameter is attached to the item which is updated each time it is accessed so that the higher the parameter the more likely it is to be accessed. But this is exactly the approach Sperber and Wilson eschew.

It appears that current notions of relevance are inadequate to the task of determining the relevant domains of knowledge which are updated in response to new information. Neither relevance logic nor relevance theory provide any grounds for believing that such an account is likely to be forthcoming.

7.4 Better Ontology

It might be thought that the locus of the problem for the logicist programme is the insistence that the rules encoding our common-sense knowledge adopt our everyday ontology of tables, chairs and so on—*i.e.* the ontology implicit in folk-psychological propositional attitude ascriptions. Perhaps

according to some more fine-grained ontology, what appear to be defeasible rules can be reconstructed as exceptionless generalisations, thus obviating the need for non-monotonic reasoning. A search for deterministic rules underlying apparently non-deterministic phenomena is analogous to Einstein's deterministic 'hidden variable' interpretation of quantum mechanics. However, the very error prone nature of most human perception, inference and action appears to militate against the possibility that people actually employ such an ontology. Any explanation of cognition must surely account for making mistakes, changing our minds, reviewing our beliefs in the light of new information *etc.* It appears necessary to *explain* the defeasibility of human inference, and impossible to *explain it away*.

Further, to retreat to the postulation of an alternative ontology, which does not correspond to everyday objects and relations, amounts to giving up point 4 in our characterisation of logicist cognitive science. This may not be a concern to many working on formalising common-sense knowledge. For example, Hayes (1984b) attempts to formalise our implicit understanding of the behaviour of liquids by postulating representational primitives which do not correspond one to one with the everyday concepts provided by pre-theoretic intuitions. Such primitives must be postulated in any case to handle inferential processes in specific cognitive domains: as we noted above, a variety of linguistic representations appear to be implicated in language understanding; a complex range of representations is computed in perceptual processes; and so on.

The rejection of everyday properties and relations as the basis for internal representation does, however, constitute a significant retreat for the logicist position of Fodor and Pylyshyn (1988). Fodor (*e.g.* 1987) and Pylyshyn (1984) argue that scientific cognitive explanation must be founded on folk-psychological explanation. Specifically, they advocate the Representational Theory of Mind according to which to have a propositional attitude is to stand in a certain relation (the relation of believing, desiring or whatever) to a mental representation. The content of this mental representation is the object of the propositional attitude. Since the contents of propositional attitudes are described in natural language, the interpretation of the corresponding mental representations must be at the level of everyday objects and relations. No everyday properties and relations, no theory of propositional attitudes.

7.5 Domain–specificity

We have observed that as the size of the knowledge-base increases, the complexity of consistency-checking becomes unacceptable, and non-monotonic logics over that knowledge-base become infeasible. If, however, knowledge can be encoded in small, isolated sets of domain-specific axioms, perhaps the complexity of consistency checking may be kept within acceptable bounds. However, it is not sufficient to maintain consistency

within domains; consistency must be maintained *between* domains, on the proof-theoretic story. As we noted above, Fodor (1983) is committed to the view that common-sense inference is precisely a domain which does not admit such modularisation. In particular, he notes common-sense inference is *isotropic*. That is, any piece of knowledge may be made to bear on any other—there are no proscribed boundaries over which inferential processes cannot operate.

We have already seen that general principles like relevance fail to provide a basis for the compartmentalisation of knowledge into specific domains. Such general principles are required since otherwise it is opaque as to how such compartmentalisation is achieved, other than by fiat, from the flux of information which an organism receives in interacting with its environment. However, let us suppose, counterfactually, that such compartmentalisation can be achieved. We now present an example which demonstrates the soundness of Fodor's intuition that domain specificity can not be the rule in knowledge based systems (on the assumption that such demonstration may still be required).

On any reasonable principles of modularisation, seismographic knowledge is unlikely to be included in the domain-specific knowledge that allows Egon to predict that his egg will be medium boiled in five minutes. However, suppose Egon is boiling his egg at the seismographic station monitoring the San Andreas fault. Egon notices the meter reading shoot off the scale. He infers that the building will be knocked flat in a few seconds and rushes out of the door. He subsequently realises that his egg will not be ready as usual, since the pan is unlikely to remain on the stove. So his knowledge of seismology seems to be implicated in explaining his expectations about eating eggs. It could reasonably be countered that Egon might not, in practice, make this inference in such a desperate situation. However, if knowledge were organised into completely isolated, domain-specific modules, he could not, *in principle*, make this inference, which seems counterintuitive. Insofar as inference *can* be based on premises from more than one knowledge domain, the axioms of each must be mutually consistent. So, since any knowledge domain *may* bear on any other, the global consistency of the entire knowledge-base must be maintained, according to the proof-theoretic view. So appeals to domain-specificity cannot alleviate the problems of consistency checking for the proof-theoretic view of common-sense reasoning.

7.6 Explicit and Implicit Inference

One line of retreat for the logicist is to grant that proof-theory does not account for defeasible inference in common-sense reasoning, language processing, perception and the like. Perhaps though, it *can* account for our explicit, conscious reasoning abilities. In explicit reasoning, only a very few premises can be entertained (Wason & Johnson-Laird, 1972; Evans, 1982; Johnson-Laird, 1983). Since in these cases the input length n is small,

the onset of the combinatorial explosion of consistency checking may be avoided. Indeed, some generally intractable exponential-time algorithms can out-perform generally tractable polynomial-time algorithms for small n. The conjecture that this is so might be supported by the fact that, given more than about three premises, in an explicit reasoning task, reasoning performance degrades catastrophically (Johnson-Laird, 1983, pp. 44–5).

There are two reasons why even this retreat may be untenable. Firstly, performance in explicit deductive reasoning tasks is extremely poor whatever the number of premises involved. This is, at least *prima facie*, puzzling if the basis of our inferential performance is proof theory (Oaksford, Chater & Stenning, 1990). Secondly, performance even on explicit deductive reasoning tasks appears to be infected by the effects of stored world knowledge (Cheng & Holyoak, 1985; Cheng, Holyoak, Nisbett & Oliver, 1986; Oaksford, 1988; Byrne, 1989; Cosmides, 1989; Evans, 1989). To model the interaction between the small number of explicitly given premisses and the huge amount of implicit world knowledge appears to require (i) that n is, after all, very large; and (ii) that a non-monotonic logic may be required to model the influence of defeasible world knowledge on deductive reasoning performance.

7.7 Can Probabilities Help?

In discussing 'relevance' we mentioned that Sperber and Wilson (1986) explicitly reject the idea of attaching subjective probabilities to propositions in memory. We now consider the possibility that so doing may go some way to resolving the problems we have raised. There seem to be two major ways in which probabilities may help. First, the defeasibility of rules which embody peoples' world knowledge need not be encoded as default rules, rather it could be conceded that all such rules are treated as probabilistic. This appears to satisfy the tractability criterion, since consistency checking over the whole data-base would no longer be required. However, we have also observed, in discussing non-monotonic logics and relevance, that defeasible inference regimes are required to solve the more general problem of which rules are to be updated in response to new information, i.e. the ubiquitous frame problem in AI. Treating *all* the rules which encode encyclopaedic world knowledge as probabilistic does not resolve the problem of *which* rules apply in a given context. Further problems, which relate to the completeness* criterion, also arise for this putative probabilistic solution.

Perhaps the most principled way of assigning probabilities to rules is given by Adams' (1966, 1975) *probability semantics*. Logically, rules are conditional statements, and hence concern centres on how probabilities should attach to conditionals. Adams suggests that the probability of a conditional, P(if ϕ, then ψ), is the *conditional probability* of ψ given ϕ. However, this proposal is subject to a well known triviality result due to Lewis (1976), the upshot of which is that such an assignment of probabilit-

ies is only possible on the assumption that ϕ and ψ are not themselves logically complex conditional statements. This result seriously restricts the scope of Adams' theory in representing very simple reasoning problems (Veltman, 1985, p. 40) and thus strongly suggests that such a proposal fails to meet the completeness* criterion.

Further grounds to believe that probabilistic rules will fail to be complete* are suggested by examples where probabilistic rules appear to act as blocks to further empirical inquiry. An example due to Alice ter Meulen (1986) can be adapted to illustrate the problem. She poses the question of what response is appropriate on encountering a complaisant donkey, given you believe that *all donkeys are stubborn*. The latter rule can be represented as a conditional statement and hence we may ask how it is to be revised in the light of this putative counter-example. Assuming that one such donkey is not taken to falsify the rule outright, the probabilistic suggestion would appear to be that a minor adjustment in the conditional probability assigned to the rule is required. Having made the adjustment, you can proceed on your way. However, surely it is at least possible that you want to *inquire* into why this particular donkey does not conform to your aforementioned belief that all donkeys are stubborn. On so inquiring, you may discover that the animal was circus trained, and hence you would be advised to encode the information that all donkeys are stubborn *except circus trained donkeys*. Such an adjustment would surely better equip you to draw appropriate inferences on next encountering a donkey at the circus, than the minor adjustment to the conditional probabilities suggested by the probabilistic alternative. It would appear that *only if* no such default information could be found (*i.e.* no hidden variables can be discovered, see above) would the probabilistic alternative be necessitated. Again it would appear necessary to *explain* the defeasibility of human inference, and impossible to *explain it away*.

A second proposal concerning how probabilities may help involves employing probabilities to determine which rules apply in which contexts. Thus rules are not soft and probabilistic but hard and logical. However, which rules apply is given by their probabilities of applying in a given context. This proposal is indistinguishable from Relevance Theory, except in the explicit use of probabilities which at least avoids the opaqueness of *confirmation strengths* in Sperber and Wilson (1986). If rules have probabilities assigned indicating their likelihood of applying in a given context, then at least two things are required: (i) a probability assignment to each rule *for each possible context*, (ii) a means of determining the current context. However, as with relevance theory, (ii) is just a restatement of the current problem. If the current context could be determined, then the problem we have invoked probabilities to resolve would not arise. Moreover, (i) requires that each rule has a probability assigned *for every possible context*. Not only is this an impossible requirement—the range of possible contexts is simply not known—the spectre of intractability must again loom very large. In, for example, computational accounts of *abductive reasoning* in

medical diagnosis, which employ Bayesian inference, the number of stored *a priori* and conditional probabilities increases explosively with the number of diseases and symptoms (Charniak & McDermott, 1985). Yet such a knowledge base must be regarded as trivial in comparison to the whole of world knowledge. In sum, there are strong grounds for believing that the probabilistic approach will be subject to intractability problems and for believing that such approaches violate the completeness* criterion.

7.8 Parallelism

From the discussion so far, it might be thought that the computational intractability of non-monotonic inference applies only to serial machines, in which computational operations must be executed one after the other. Perhaps an appeal to the parallelism of the brain may alleviate the problem of computational intractability. However, at best, appeals to parallelism can only reduce the time-complexity of a computationally explosive algorithm by a constant factor. All that this can do is slightly delay the onset of the computational infeasibility. Given that the proof-theoretic view of mind requires consistency checking over a data-base which encodes the whole of an individual's common-sense knowledge (so that n is, presumably, *very* large), the minor gains induced by appeal to parallelism are unlikely to be significant.

7.9 Semantic Methods of Proof

Within the psychology of reasoning, there is an important debate about whether or not human *deductive* reasoning is mediated by proof theoretic methods (Piaget, 1953; Henle, 1962; Braine, 1978), or by semantic methods of proof, such as mental models (Johnson-Laird, 1983). Carrying this debate over to non-demonstrative inference, it might be thought that such semantic methods may provide an alternative to the standard proof-theoretic approach. However, such semantic methods of proof *work by* consistency checking. The validity of an inference from $A_1, A_2 \ldots A_n$ to a conclusion C, is established by attempting to show that $A_1, A_2 \ldots A_n \,\&\, \neg C$ is not consistent. The consistency check is performed by systematically attempting to find a model according to which each of $A_1, A_2 \ldots A_n \,\&\, \neg C$ are true. Since consistency checking, *of whatever form*, is NP-complete, as the number of premises in the data-base increases, the computation becomes intractable, and inferences cannot be made, even in a *monotonic* logic. Although semantic methods such as mental models may elegantly account for some explicit deductive reasoning tasks, they offer no prospect of providing more tractable mechanisms for reasoning in knowledge-rich domains.

7.10 Proceduralism

McDermott (1986) argues that the failure of proof-theoretic methods in AI to adequately account for non-monotonic reasoning requires that the attempt to provide a semantics for knowledge representation formalisms must be abandoned. Yet this move amounts to abandoning the project of accounting for defeasible reasoning. If symbolic structures are assigned no semantics, then they have no representational content—that is, they are not *about* anything. Yet reasoning processes are defined in virtue of the content of the representations that they manipulate; to describe an inference as valid, justified, legitimate is to appeal to the interpretation of the symbolic structures. For example, the inference from A and A → B to B is *valid* since if A and A → B are both *true*, then B must be *true*. Yet uninterpreted formulae, which are all that the proceduralist can countenance, cannot be true or false.

The only retreat is to appeal to some form of Functional Role Semantics (see Block, 1986, for a review and references). That is, the idea that symbols can acquire meanings via their intrinsic relations to other symbols. This idea is usually illustrated (see for example, Lloyd, 1989, pp. 24–5) by an analogy with learning the meaning of a term either in a foreign language or in an unfamiliar idiolect of a speaker's own language. A previously unencountered term may be acquired and used appropriately simply by observing its relations to other words, and its grammatical contexts of use. A speaker may finally become competent enough to use the term appropriately to utter truths *without ever having learned the precise denotation of the term, i.e.* without access to the full semantic content of the symbol. However, it is generally agreed that this story can not work for all the terms of a language (Lloyd, 1989). At least some, more likely the majority, of the terms of a speaker's language must be such that the speaker has access to their full semantic content. Without such access no sense could be attached to talk of 'using a term appropriately to utter truths'. It is important to observe that Fodor himself does not believe a word of the functional role story (see, Fodor, 1987). Although this view is easily conflated with Fodor's (1980) *methodological solipsism*, there is nothing *methodological* about it, this is *solipsism* pure, simple and indefensible. Without appeal to full semantic content, cognitive science does not have a story to tell about its central explanatory concept, *i.e.* representation.

Quite generally, proceduralism abandons all notions of reasoning and inference, be they deductive, inductive, eductive or abductive. *Very* generally, it is hard to imagine what a cognitive science (logicist or otherwise) could look like, without the notion of representation.

8. Conclusions

We have argued that the plausibility of logicist cognitive science depends on its ability to provide a proof-theoretic account of defeasible inference which is implicated in almost every area of cognitive activity. We assessed the practical attempt in AI to carry out this proof-theoretic programme using non-monotonic logics, and noted (1) that such logics are able to draw only unacceptably weak disjunctive conclusions; and (2) that the theorem-proving algorithms over such logics are computationally intractable due to their reliance on the NP-complete problem of consistency checking. We drew the conclusion that the programme of logicist cognitive science is infeasible, and replied to a number of plausible objections to this conclusion.

If logicist cognitive science constitutes an inappropriate framework in which to model cognition, the question arises of what alternative approach can be provided, which maintains both semantic interpretability and computational tractability. In discussing the central dogmas of logicist cognitive science, we repeatedly urged that the range of computational systems available is far from exhausted by the traditional symbolic approach. Nevertheless, it is beyond dispute that this is the approach which has been most thoroughly investigated, in part because of its early promise in providing a physicalist grounding for human cognitive processes. However, in virtue of this fact, it is the approach about which most is known relative to its abilities to handle cognitive phenomena. From the issues raised here concerning the defeasibility of human cognitive processes, it is clear that the conclusion of these investigations is that classical logicist cognitive science is inadequate. Therefore, it may well be time to explore the space of possible computational schemes for more adequate, albeit as yet less well understood, alternatives.

In this regard, recent work on distributed systems such as neural networks (*e.g.* Rumelhart & McClelland, 1986; McClelland & Rumelhart, 1986) and classifier systems (Holland, Holyoak, Nisbett & Thagard, 1986), may perhaps constitute the beginnings of an alternative approach to mechanisms which deal with defeasible inference (Shastri, 1985; Derthick, 1987; Chater & Oaksford, 1990). Both Shastri (1985) and Derthick (1987, 1988) provide efficient implementations of algorithms for Bayesian inference and default reasoning respectively, which exploit connectionist systems. However, both implementations are hand-wired and thus do not exploit the principle advantage of connectionist systems, *i.e.* their ability to learn. Connectionist learning is notoriously slow, and thus our suggestion that such systems may aid in overcoming the objections to the logicist programme we raise in this paper may seem suspect. Complexity results for connectionist learning algorithms are as bad (in fact usually worse) than the non-monotonic systems we criticise. However, like is not being compared with like. In the case of non-monotonic systems the complexity of inference *not* learning was under discussion: these systems do not possess learning mechanisms. With regard to inference in Connectionist systems, once a network has learned, it draws inferences as rapidly as it takes to propagate activity from input to output. This pattern of complexity mirrors the human case whereas that of non-monotonic reasoning systems does

not. Human learning is a slow process, but once some piece of knowledge is in place inference over it is effortless. Connectionist systems appear to display precisely the same complexity profile.

Cognitive Neurocomputation Unit
Department of Psychology
University of Wales
Bangor, Gwynedd LL57 2DG
Wales

Department of Psychology
University of Edinburgh
7 George Square
Edinburgh
EH8 9JZ

References

Adams, E. 1966: Probability and the Logic of Conditionals. In J. Hintikka and P. Suppes (eds.), *Aspects of Inductive Logic*, Amsterdam: North Holland.

Adams, E. 1975: *The Logic of Conditionals: An Application of Probability to Deductive Logic*. Dordrecht: Reidel.

Anderson, A.R. and Belnap, N.D. 1975: *Entailment: The Logic of Relevance and Necessity*, Vol. 1. Princeton: Princeton University Press.

Block, N. 1986: Advertisement for a Semantics for Psychology. *Midwest Studies in Philosophy*, Volume 10, 615–678.

Boolos, G. and Jeffrey, R. 1980: *Computability and Logic*. 2nd Edition, Cambridge: Cambridge University Press.

Braine, M.D.S. 1978: On the Relationship Between the Natural Logic of Reasoning and Standard Logic. *Psychological Review*, 85, 1–21.

Bransford, J.D., Barclay, J.R. and Franks, J.J. 1972: Sentence Memory: A Constructive versus Interpretive Approach. *Cognitive Psychology*, 3, 193–209.

Bransford, J.D. and Johnson, M. 1972: Contextual Prerequisites for Understanding: Some Investigations of Comprehensions and Recall. *Journal of Verbal Learning and Verbal Behaviour*, 11, 717–726.

Bransford, J.D. and Johnson, M.K. 1973: Considerations of Some Problems of Comprehension. In W.G. Chase (ed.), *Visual Information Processing*, New York: Academic Press, 389–392.

Bransford, J.D. and McCarrell, N.S. 1975: A Sketch of a Cognitive Approach to Comprehension: Some Thoughts on What It Means to Comprehend. In W.B. Weimer and D.S. Palermo (eds.), *Cognition and Symbolic Processes*, Hillsdale, N.J.: Lawrence Erlbaum Associates, 189–229.

Byrne, R.M.J. 1989: Suppressing Valid Inferences With Conditionals. *Cognition*, 31, 1–21.

Carnap, R. 1923: Über die Aufgabe der Physik und die Anwendung des Grundsatzes der Einfachheit. *Kant-Studien*, 28, 90–107.

Carnap, R. 1950: *Logical Foundations of Probability*. Chicago, Illinois: University of Chicago Press.

Chater, N. and Oaksford, M. 1990: Autonomy, Implementation and Cognitive Architecture: A Reply to Fodor and Pylyshyn. *Cognition*, 34, 93–107.

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

Cheng, P.W. and Holyoak, K.J. 1985: Pragmatic Reasoning Schemas. *Cognitive Psychology*, 17, 391–416.

Cheng, P.W., Holyoak, K.J., Nisbett, R.E. and Oliver, L.M. 1986: Pragmatic versus Syntactic Approaches to Training Deductive Reasoning. *Cognitive Psychology*, 18, 293–328.

Clark, H.H. 1977: Bridging. In P.N. Johnson-Laird and P.C. Wason (eds.), *Thinking: Readings in Cognitive Science*, Cambridge: Cambridge University Press, 411–420.

Clark, K.L. (1978) Negation as Failure. In *Logic and Databases*, New York: Plenum Press, 293–322.

Clocksin, W.F. and Mellish, C.S. 1984: *Programming in Prolog*. Second edition, Berlin: Springer-Verlag.

Cook, S. 1971: The Complexity of Theorem Proving Procedures. In *The Third Annual Symposium on the Theory of Computing*, New York, NY, 151–158.

Cosmides, L. 1989: The Logic of Social Exchange: Has Natural Selection Shaped How Humans Reason? Studies With the Wason Selection Task. *Cognition*, 31, 187–276.

Derthick, M. 1987: A Connectionist Architecture for Representing and Reasoning About Structured Knowledge. Research Report No. CMU-BOLTZ-29, Department of Computer Science, Carnegie Mellon University, Pittsburgh.

Derthick, M. 1988: Mundane Reasoning by Parallel Constraint Satisfaction. Research Report No. CMU-CS-88-182, Department of Computer Science, Carnegie Mellon University, Pittsburgh.

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

Evans, J. St. B.T. 1989: *Bias in Human Reasoning: Causes and Consequences*. London: Lawrence Erlbaum Associates.

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

Fodor, J.A. 1980: Methodological Solipsism Considered as a Research Strategy in Cognitive Psychology. *Behavioural and Brain Sciences*, 3, 63–109.

Fodor, J.A. 1983: *The Modularity of Mind*. Cambridge, MA.: MIT Press.

Fodor, J.A. 1987: *Psychosemantics: The Problem of Meaning in the Philosophy of Mind*. Cambridge, MA.: MIT Press.

Fodor, J.A., Bever, T.G. and Garrett, M.F. 1974: *The Psychology of Language*. New York: McGraw Hill.

Fodor, J.A. and Pylyshyn, Z.W. 1981: How Direct is Visual Perception? Some Reflections on Gibson's 'Ecological Approach'. *Cognition*, 9, 139–196.

Fodor, J.A. and Pylyshyn, Z.W. 1988: Connectionism and Cognitive Architecture: A Critical Analysis. *Cognition*, 28, 3–71.

Goodman, N. 1983: *Fact, Fiction and Forecast*. Fourth Edition, Cambridge, MA.: Harvard University Press.

Gregory, R.L. 1977: *Eye and Brain*. Third Edition, London: Weidenfeld and Nicolson.

Haack, S. 1978: *Philosophy of Logics*. Cambridge: Cambridge University Press.

Hanks, S. and McDermott, D. 1985: Default Reasoning, Nonmonotonic Logics,

and the Frame Problem. *Proceedings of the American Association for Artificial Intelligence*. Philadelphia, PA.
Hanks, S. and McDermott, D. 1986: Temporal Reasoning and Default Logics. Yale University, Computer Science Technical Report, No. 430.
Hayes, P. 1978: The Naive Physics Manifesto, In D. Michie (ed.), *Expert Systems in the Microelectronic Age*. Edinburgh, Scotland: Edinburgh University Press.
Hayes, P. 1984a: The Second Naive Physics Manifesto. In J. Hobbs (ed.), *Formal Theories of the Commonsense World*, Hillsdale, N.J.: Ablex.
Hayes, P. 1984b: Liquids. In J. Hobbs (ed.), *Formal Theories of the Commonsense World*, Hillsdale, N.J.: Ablex.
Hempel, C. 1952: *Fundamentals of Concept Formation in Empirical Science*. Chicago, Illinois: University of Chicago Press.
Hempel, C. 1965: *Aspects of Scientific Explanation and Other Essays in the Philosophy of Science*. New York: Free Press.
Henle, M. 1962: On the Relation Between Logic and Thinking. *Psychological Review*, 69, 366–378.
Hogger, C.J. 1984: *An Introduction to Logic Programming*. London: Academic Press.
Holland, J.H., Holyoak, K.J., Nisbett, R.E. and Thagard, P.R. 1986: *Induction: Processes of Inference, Learning and Discovery*. Cambridge, MA.: MIT Press.
Horowitz, E. and Sahni, S. 1978: *Fundamentals of Computer Algorithms*, Rockville, Maryland: Computer Science Press, Inc.
Israel, D.J. 1980: What's Wrong With Non-monotonic Logic? In *Proceedings of AAAI-80*, 1980, 99–101.
Johnson-Laird, P.N. 1983: *Mental Models*, Cambridge: Cambridge University Press.
Lakatos, I. 1970: Falsification and the Methodology of Scientific Research Programmes. In I. Lakatos and A. Musgrave (eds.), *Criticism and the Growth of Knowledge*, Cambridge: Cambridge University Press, 91–196.
Levesque, H.J. 1988: Logic and the Complexity of Reasoning. *Journal of Philosophical Logic*, 17, 355–89.
Lewis, D. 1976: Probabilities of Conditionals and Conditional Probabilities. *Philosophical Review*, 85, 297–315.
Lloyd, D.E. 1989: *Simple Minds*. Cambridge, MA.: MIT Press.
Loui, R.P. 1986: Defeat Among Arguments: A System of Defeasible Inference. Technical Report No. 190, Department of Computer Science, Rochester University.
Loui, R.P. 1987: Response to Hanks and McDermott: Temporal Evolution of Beliefs and Beliefs About Temporal Evolution. *Cognitive Science*, 11, 283–97.
Marr, D. 1982: *Vision*. San Francisco: W.H. Freeman and 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.
McDermott, D. 1982: Non-monotonic Logic II: Non-monotonic Modal Theories; *JACM*, 29(1), 33–57.
McDermott, D. 1986: A Critique of Pure Reason. Technical Report, Department of Computer Science, Yale University, June, 1986.
McDermott, D. and Doyle, J. 1980: Non-monotonic Logic I. *Artificial Intelligence*, 13, 41–72.

McClelland, J.L. and Rumelhart, D.E. 1986: *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, Volume 2: *Psychological and Biological Processes*. Cambridge, MA.: MIT Press.
Medin, D.L. and Schaffer, M.M. 1978: Context Theory of Classification Learning. *Psychological Review*, 85, 201–238.
ter Meulen, A. 1986: Generic Information, Conditional Contexts and Constraints. In E.C. Traugott, A. ter Meulen, J. Snitzer Reilly and L.A. Ferguson (eds.), *On Conditionals*, Cambridge: Cambridge University Press.
Minsky, M. 1975: Frame-system Theory. In R. Schank and B.L. Nash-Webber (eds.), *Theoretical Issues in Natural Language Processing*, Cambridge, MA. June 10–13, 1975.
Nosofsky, R.M. 1986: Attention, Similarity and the Identification–Categorisation Relationship. *Journal of Experimental Psychology: General*, 115, 39–57.
Nute, D. 1985: A Non-monotonic Logic Based on Conditional Logic. Working Paper, Advanced Computational Methods Centre, University of Georgia, Athens, GA.
Nute, D. 1986: A Logic for Defeasible Reasoning. Research Report No. 01–0013, Advanced Computational Methods Centre, University of Georgia, Athens, GA.
Oaksford, M. 1988: Cognition and Inquiry: The Pragmatics of Conditional Reasoning. PhD Thesis, Centre for Cognitive Science, University of Edinburgh.
Oaksford, M., Chater, N. and Stenning, K. 1990: Connectionism, Classical Cognitive Science and Experimental Psychology. *AI and Society*, 4, 73–90.
Peirce, C.S. 1931–1958: *Collected Papers* (Eight Volumes). C. Hartshorne, P. Weiss and A. Burks (eds.), Cambridge, MA.: Harvard University Press.
Piaget, J. 1953: *Logic and Psychology*. Manchester: University of Manchester Press.
Poole, D. 1985: On the Comparison of Theories: Preferring the Most Specific Explanation. In *Proceedings of the International Joint Conference on Artificial Intelligence*, Los Angeles, CA.
Popper, K.R. 1959: *The Logic of Scientific Discovery*. London: Hutchinson.
Pylyshyn, Z.W. 1973: What the Mind's Eye Tells the Mind's Brain: A Critique of Mental Imagery. *Psychological Bulletin*, 80, 1–24.
Pylyshyn, Z.W. 1984: *Computation and Cognition: Toward a Foundation for Cognitive Science*. Montgomery, Vermont: Bradford.
Reiter, R. 1980: A Logic for Default Reasoning. *Artificial Intelligence*, 13, 81–132.
Reiter, R. 1985: On Reasoning by Default. In R. Brachman and H. Levesque (eds.), *Readings in Knowledge Representation*, Los Altos, California: Morgan Kaufman (originally 1978).
Rosch, E. 1973: On the Internal Structure of Perceptual and Semantic Categories. In T. Moore (ed.), *Cognitive Development and the Acquisition of Language*, New York: Academic Press.
Rosch, E. 1975: Cognitive Representation of Semantic Categories. *Journal of Experimental Psychology: General*, 104, 192–233.
Rumelhart, D.E. and McClelland, J.L. 1986: *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, Volume 1: *Foundations*. Cambridge, MA.: MIT Press.
Shastri, L. 1985: Evidential Reasoning in Semantic Networks: A Formal Theory and its Parallel Implementation. TR166, Department of Computer Science,

University of Rochester, September, 1985.
Shoam, Y. 1986: Chronological Ignorance: Time, Non-monotonicity, Necessity and Causal Theories. In *Proceedings of the Association of Artificial Intelligence*, Philadelphia, PA.
Shoam, Y. 1987: *Reasoning About Change*. Cambridge, MA.: MIT Press.
Shoam, Y. 1988: Efficient Reasoning About Rich Temporal Domains, *Journal of Philosophical Logic*, 17, 443–474.
Sperber, D. and Wilson, D. 1986: *Relevance: Communication and Cognition*. Oxford: Basil Blackwell.
Stenning, K. and Oaksford, M. 1989: Choosing Computational Architectures for Text Processing. Technical Report No. EUCCS/RP–28, Centre for Cognitive Science, University of Edinburgh, April, 1989.
Suppe, F. 1977: *The Structure of Scientific Theories*, Second Edition, Urbana: University of Illinois Press.
Veltman, F. 1985: Logics for Conditionals. PhD. Thesis, Faculteit der Wiskunde en Naturwetenschappen, University of Amsterdam.
Wason, P.C. and Johnson-Laird, P.N. 1972: *The Psychology of Reasoning: Structure and Content*. Cambridge, MA.: Harvard University Press.