Mind & Language
Vol. 8 No. 1 Spring 1993

Article

Logicism, Mental Models and Everyday Reasoning: Reply to Garnham

NICK CHATER AND MIKE OAKSFORD

Alan Garnham (this issue, pp. 49–71) has provided a lucid and thoughtful challenge to our arguments against Logicist cognitive science (Oaksford & Chater, 1991). He argues that many of our arguments are misdirected or fallacious, and that we draw entirely the wrong moral from the comparison of human reasoners and logic-based artificial reasoning systems. Some of Garnham's objections are due to a misreading of our argument against Logicism. We first reiterate the structure of that argument and then show that many of Garnham's points are best read as supportive of our conclusions though critical of our presentation. Garnham's central point, that mental models theory supplies a distinct, and distinctly more promising, alternative to Logicist cognitive science, requires a more substantial treatment, however. We argue that mental models provides no defence against the twin difficulties of intractability and incompleteness* that we raised for Logicism.

1. The Structure of Our Argument

Our argument ran as follows. Firstly, we characterized Logicist cognitive science, the theoretical view of the nature of cognitive science expounded by Fodor and Pylyshyn (*e.g.* Fodor, 1975; Pylyshyn, 1984; Fodor & Pylyshyn, 1988). Roughly, Logicism is the view that cognitive processes are

Address for correspondence: Mike Oaksford, Department of Psychology, University of Wales, Bangor, Gwynedd LL57 2DG, UK.
Email: pss027@bangor.ac.uk.

proof-theoretic operations over internal logical formulae which can be interpreted in terms of our everyday ontology of tables, chairs and so on. We took this view as a definite target at which to aim our arguments, rather than as representative of cognitive scientists at large; the degree to which our arguments carry over to variants of Logicism is deferred until later in the paper.

Secondly, we considered how such explanation might fare as an account of cognitive processes which involve knowledge-rich defeasible inference (see, for example, footnote 1 of our paper). The central processes (Fodor, 1983) involved in common-sense reasoning are paradigm examples of knowledge-rich processing. To the extent that aspects of perception, language processing and so on are also knowledge-rich, the same problems should apply.

We noted that the central processes involved in what may variously be thought of as belief revision, common-sense inference or everyday reasoning are a species of inference to the best explanation (Fodor, 1983). That is, given certain information, the reasoner must infer what fits best with, what best explains and is explained by, that information. Inference to the best explanation is notoriously difficult to capture within the framework of deductive logic. For one thing, standard deductive validity entails that if the premises of an argument are true, then the conclusion must certainly also be true. This means that standard deductive logic is *monotonic*: if a conclusion follows deductively from a set of premises, it will follow from the conjunction of that set of premises with any other additional information. Yet in inference to the best explanation a hypothesis that seemed plausible in the light of partial evidence will often seem implausible in the light of a fuller picture. That is, such inference is invariably tentative rather than certain and will be *nonmonotonic*.

This mismatch poses a serious problem for Logicism: if cognitive processes are proof theoretic, and proof theory standardly can only handle monotonic deductive reasoning, how can the non-demonstrative inferences which appear to be cognitively ubiquitous be explained? We noted that this dilemma had a historical correlate in the unsuccessful attempts of the logical positivists to cast inference to the best explanation in science in a deductive mould, and suggested that a Logicist account of central processes would be likely to fare no better. The argument could have stopped here: to the extent that common-sense and scientific inference are analogous, it should be equally easy (or difficult) to model either by proof-theoretic methods. And it is universally acknowledged in the philosophy of science that scientific inference cannot be understood in this way (Goodman, 1983; see also, Holland, Holyoak, Nisbett & Thagard, 1986; Thagard, 1988). Rather than rely solely on this argument by analogy, we turned to a practical test of the feasibility of Logicism: the attempt to model everyday reasoning within artificial intelligence.

Third, then, we considered Logicist work on building computational models of aspects of common-sense inference. The volume of such work

is vast, and the range of techniques employed is also great (see, for example, the collection edited by Ginsberg, 1987). Rather than attempt a survey, we focussed on a particular approach, which is closest to the spirit of Logicism, is dominant within artificial intelligence, and to which other approaches are very intimately related (Hanks & McDermott, 1985, 1986; Shoam, 1987, 1988). This approach involves developing *non-monotonic* logics, in which the addition of premises can lead to the withdrawal of conclusions, to account for the revisability of everyday reasoning. Thus, in principle at least, proof theory over non-monotonic logics may be able to reconcile Logicism with the defeasible character of inference in central processes. We then raised two serious and apparently fatal problems for the enterprise. Firstly, non-monotonic logics are generally not able to capture plausible but revisable everyday inferences. The conclusions licensed by such logics are, in general, irremediably weak, often to the point of total vacuity. Thus the attempt to model common-sense using non-monotonic logics has not bridged the gap between proof-theoretic methods and inference to the best explanation, but simply illustrated how great that gap is. Secondly, even if non-monotonic logics were able to model everyday inferences in principle, they would still be unviable since proof methods for such logics are radically computationally intractable. In sum, the attempt to fit apparently non-deductive common-sense reasoning into a deductive framework fails because it doesn't specify the right answers, and in practice it is so intractable that it doesn't give any answers at all. We concluded that these considerations undermine the plausibility of Logicism as a model of central cognitive processes.

The fourth step in our argument was to consider possible replies and objections. The thrust of many of the objections was that if the unnecessarily tight constraints of the Logicist position are loosened, our arguments no longer apply. Variants that we considered included: using heuristics to supplement purely proof-theoretic operations, abandoning proof theory altogether and using entirely procedural symbolic methods and denying that the internal language can be interpreted in terms of our common-sense ontology of tables and chairs. Thus, in this section, the question of how widely the arguments against the rather specific target of Logicism apply to nearby positions in cognitive science was addressed. Among the neighbours of Logicism which we considered was the use of semantic or model-based, rather than syntactic, methods of proof. Garnham argues that this dismissal was not compelling, and that such methods do not succumb to our arguments, and we shall discuss this proposal extensively below. The upshot of our discussion was that the arguments against the specific target of Logicism apply very much more widely; they hit equally forcefully at positions which respect the spirit, but not the letter, of Logicism.

2. *Have We Been Misconstrued?*

In the light of this outline (and indeed, in the light of the original text, where this structure is perhaps less clearly highlighted) many of Garnham's

points seem somewhat tangential. So, for example, Garnham's Section 2 suggests that Marr's work on vision provides an existence proof of the possibility of Logicist cognitive science. Yet Marr's work is certainly not Logicist: the operations that Marr discusses are not proof-theoretic and the internal representations Marr discusses cannot be interpreted in terms of our common-sense ontology. Moreover, Marr's work does not concern knowledge-rich processes—precisely those processes with which we are concerned. Indeed, Marr (*e.g.* 1982) is concerned to avoid knowledge-rich processes as far as possible, precisely because such processes are so little understood. So while we entirely agree with Garnham that Marr's work is an object lesson in cognitive science, we do not see this as bearing on the argument against Logicism.

Similarly, Section 4 of Garnham's discussion, which provides a detailed analysis of each of the tenets of Logicism, does not appear to be at variance with our position. In each case, the thrust of the discussion is that these claims, while acceptable to Fodor and Pylyshyn, would not necessarily be common ground in the cognitive sciences more widely. The implication is that even if our arguments against Logicist cognitive science (on the narrow Fodor/Pylyshyn reading) are valid, these arguments may not generalize to other accounts in the same spirit. Certainly, our arguments do not *necessarily* generalize. However, we argued extensively in the ***Objections and Replies*** section that they would appear to generalise in fact: the numerous variations on Logicism we considered appeared to do nothing to deflect these arguments.

The range of theoretical positions which deviate in one way or another from Fodor and Pylyshyn's position is, as Garnham amply illustrates, very broad indeed. Rather than attempt to set up an all inclusive characterization of accounts of central processes in cognitive science, we picked the most specific, best worked out and most influential account as our primary target. We then considered piecemeal whether or not variations on the strict Logicist position would be of any help. So while we agree with Garnham that the tenets of Logicism are not by any means universally accepted, this point seems to be compatible with, rather than inimical to, the conclusions of our paper.

Similarly it is not clear that Section 5 of Garnham's reply stands against our arguments against Logicism. Garnham suggests that, in replying to possible objections, we conflate the two problems we identify for non-monotonic logics: that they do not license inferences strong enough to capture everyday reasoning (the constraint that we called 'completeness*' and which Garnham calls 'adequacy') and tractability considerations. Probably, as Garnham suggests (Section 5, p. 55), it would have been helpful to explicitly label which of these problems each of the possible patches to Logicism addressed.

Nonetheless, we are not sure that there was really much room for confusion between the tractability and completeness* issues in our original ***Objections and Replies*** section. Since each of these problems was dealt

with in a separate section in the original argument, and since we stressed that completeness* and tractability pose independent problems for Logicist accounts, it is implicit that a successful objection to our arguments must show how *both* of these difficulties can be overcome. In practice, the objections that we considered can generally only handle one of these objections *at best*, and we were concerned to show that even such minimal inroads could not be sustained.

Garnham goes on to argue that our discussion of the tractability of non-monotonic logics is beside the point if adequacy criteria cannot be met. 'There is no point in worrying about the computational properties of a system, if that system can be rejected as a model of everyday reasoning on the grounds that it is irredeemably inadequate.' (Section 5, p. 55). '. . . if nonmonotonic logics don't capture everyday reasoning, why try to draw conclusions about the nature of cognitive science on the assumption that its models of everyday reasoning will be based on nonmonotonic logic? An obvious tactic would be to look elsewhere' (Section 5, p. 55).

It is difficult to disagree with these sentiments. We too suspect that meeting completeness* (adequacy) poses insuperable problems for Logicist accounts; and hence that the conclusion stands even without the tractability considerations (actually Garnham thinks that our conclusions concerning tractability, particularly in relation to human inference, are wrong-headed in a rather different way, which we consider later). On the other hand, not all readers may be as convinced as Garnham by completeness* considerations, and some may find the second line of attack more compelling. Furthermore, the tractability problems of non-monotonic logics are an instructive illustration of the appalling computational tangle that results from trying to assimilate non-deductive reasoning to a deductive framework. In any case, it is clear that the only point of disagreement (if any) concerns economy of presentation and that none of these points rebut our conclusions. Regarding Garnham's additional point, that given that non-monotonic logic violates completeness*, we should look elsewhere, again we agree. As we noted above, in our ***Objections and Replies*** section we devoted considerable space to a number of possible alternatives.

It seems likely that there is also no substantial disagreement over our discussion of heuristics, although the use of the term, borrowed from the literature on knowledge representation in artificial intelligence, may indeed have puzzled some readers (Garnham, Section 5, pp. 56–8). Certainly, the term 'heuristic' is generally used to refer to a quick but fallible computational trick to shortcut a computationally expensive algorithmic computation. Accordingly, there is no possibility that heuristics can give correct answers when the algorithm does not, only that they can arrive at an answer more quickly. In the present context, appeal to heuristics in this sense could indeed only address tractability and certainly not completeness*/adequacy. The sense of heuristic with which we were working, borrowed from the knowledge representation literature in artificial intelligence, *does*, however, place the onus on heuristics embodying constraints

that allow a computational system to obtain the right (common-sense) inferences, when application of the proof theoretic approach would not do so alone (see *e.g.* Hanks & McDermott, 1985, 1986; Loui, 1987). Thus Garnham is entirely right to note, 'No wonder O & C conclude that explanatory power has been shifted from the logic to the heuristics: they are trying to make the heuristics get things *right* when the algorithmic procedure gets them wrong!' (Garnham, Section 5, p. 57).

Quite generally, the thrust of Garnham's comments, while written as if they were hostile to our position, appear to be read better as a series of points concerning how our argument might have been made more briefly, less confusingly and so on, and reveal no real points of disagreement. It is in Sections 6 and 7 that Garnham counters our arguments directly. While granting that Logicist cognitive science, strictly characterized, may fall victim to the kind of arguments that we present, he suggests that semantic methods of proof, and in particular approaches to inference within the framework of mental models, may not succumb to this line of reasoning.

3. *Semantic Methods of Proof*

The discussion of semantic methods of proof, in our section ***Objections and Replies*** in the original paper, briefly considered whether or not semantic methods of proof could address the issue of tractability. Semantic methods of proof are based on the search for a model which provides a counterexample to the inference, *i.e.* a model in which the premises are true but the conclusion is false. If such a model can be found, the inference is not valid; if there is no such model, then the inference is valid. Since the space of models which must be considered grows exponentially with the number of premises under consideration we concluded that semantic methods of proof are unpromising with respect to providing a solution to the tractability problem. Indeed, within the study of theorem proving in computer science, syntactic methods of proof are preferred as being more tractable than their semantic counterparts.

Garnham grants that semantic methods of proof are computationally intractable, but argues that when the nature of human inferential performance is properly analysed, tractability is revealed to be a pseudoproblem. He also suggests that semantic methods, and in particular the mental models framework (Johnson-Laird, 1983), may be able to address the completeness* problem: that semantic methods of proof have the potential to account for everyday inferences. For appeal to semantic methods of proof to be effective, clearly both of these claims must be upheld. We shall argue that, on the contrary, neither of them can be defended.

3.1 Semantic Methods and Tractability

Garnham provides both general and specific arguments that complexity is not the problem that we take it to be. The general argument is that the fact that an algorithm is intractable does not necessarily mean that it cannot be successfully used in practice. First we '. . . have no direct argument against the claim that proof procedures for adequate nonmonotonic logics (if there be such things) might run into problems only on problems that are never encountered in everyday life' (Garnham, Section 6, p. 60). And second our 'arguments do not generalize to model-theoretic accounts that are not directly related to failed nonmonotonic logics'.

With respect to the first point, it seems to us that the boot is securely on the other foot. It is up to the proponent of a computational scheme which is computationally intractable to explain why practical problems will not in fact arise. In the absence of any reason to suspect that this is true, there is surely every expectation that such a remarkably convenient state of affairs will not arise. Since non-monotonic logics (and related schemes) require an (intractable) consistency check *every time a plausible inference is made*, and this consistency check is performed over the *entire knowledge base* (or at best over a very large fragment of this knowledge—see the discussion of Domain Specificity in the ***Objections and Replies*** section of our original paper), it seems extremely unlikely that tractability problems can be avoided. As we noted in the original paper, the fact that no reasoning system based on a non-monotonic logic has been implemented with more than a handful of premises testifies to the drastic limitations that the problem of intractability imposes.

It is difficult to know what underlies the second point, that our arguments do not generalize to semantic methods of proof. If semantic methods of proof offer no succour with respect to tractability, as Garnham admits, it seems that generalization to semantic methods has already been granted.

The specific reasons why Garnham suspects that complexity is not a problem is that human reasoning is actually susceptible to complexity considerations. It is, after all, well known that, as the number of premises in a reasoning task increases beyond 2 or 3 reasoning performance collapses catastrophically. So, Garnham argues '. . . if a semantically-based account of human reasoning predicts that the problems become intractable, and hence impossible to solve in a reasonable amount of time, as the number of premises increases, so much the better. To the extent that it does, it accurately models human performance' (Garnham, Section 6, p. 60).

This argument seems to be entirely beside the point. What is under consideration is common-sense reasoning, rather than deductive reasoning. In deductive reasoning, to be sure, human performance is extraordinarily poor and brittle, and only very minute problems can be tackled (*e.g.* Johnson-Laird 1983, pp. 44–5). Yet this stands in direct contrast to the case of common-sense reasoning, where we appear to be able to effortlessly recruit vast amounts of knowledge in drawing plausible conclusions (indeed, the entire knowledge base may be in play, rather than two or three premises).

What conclusion should we draw from the drastic limitations on human deductive reasoning, in comparison to our facility at everyday reasoning? There are two broad answers, neither of which offer comfort to the Logicist. One possibility is that these different species of reasoning are effected by entirely different processes, one of which is very poorly developed and inefficient and one of which is remarkably powerful and fast. If this is correct, then the complexity profile of human deductive reasoning is irrelevant to the question in hand: providing a tractable and adequate account of commonsense inference.

A second, perhaps more interesting, possibility is that the same mechanism is responsible for both deductive reasoning and the inference to the best explanation involved in common-sense reasoning. If so, then the disparity in the levels of human performance between the two can best be explained by assuming that central processes are adapted to common-sense reasoning, and only co-opted into performing deductive reasoning (Oaksford, Chater & Stenning 1990; Oaksford & Chater 1992a, 1992b; Oaksford & Stenning, 1992). Consider an analogy with human locomotion. The properties of the limbs are presumably highly adapted to walking and running, at which they are very successful. The limbs are also crucially involved in walking on one's hands, to which they are not adapted, and at which performance is very poor. Structures which originally have one function can, if necessary, be co-opted to perform some other function. So, one might imagine, the mental apparatus whose function is common-sense reasoning may be co-opted to attempt to solve deductive reasoning problems, although performance would be expected to be poor. If there is a single underlying mechanism subserving common-sense and deductive reasoning, then the study of a putative underlying mechanism should presumably focus on its operation in tasks to which it is adapted, rather than in tasks for which it is not primarily designed, just as the study of locomotion focuses on walking and running rather than on more arcane ways of moving about.

If this is right, theories which are primarily constructed to model deductive reasoning performance are *prima facie* unlikely to be good candidates as theories of common-sense reasoning, just as a theory of human locomotion which focussed on hand-walking data and attempted to generalize to walking and running would be unlikely to be of value. This is, however, precisely the strategy that Garnham adopts. He considers the mental models account of deductive reasoning as a sound foundation for a model of the general case, common-sense reasoning, even though he considers that deductive and common-sense reasoning may well not be carried out by the same mechanisms. Our locomotion analogy would be no more than a straw in the wind in the absence of independent grounds for believing that mental models are not an adequate account of common-sense reasoning. It does however illustrate why it may be an unreasonable, though

not unusual (*e.g.* Johnson-Laird 1983; Johnson-Laird & Byrne 1991) expectation that mental models theory will generalize from deductive to non-deductive reasoning.

We have argued that tractability considerations are both severe and germane for theories according to which the cognitive system employs semantic, rather than syntactic, methods of proof. Thus, with regard to complexity considerations there seems to be every reason to suppose that semantic methods of proof cannot be the basis of common-sense inference, over very large bodies of information, which people so rapidly and routinely perform. As we shall now see, semantic methods of proof are equally unable to address the problem of completeness* or adequacy. Just as with syntactic methods of proof, semantic methods would give the wrong answers, if they were computationally tractable enough to give any answers at all.

*3.2 Semantic Methods and Completeness**

Is it possible that semantic methods of proof can provide the extra 'power' required to account for the strength of common-sense inferences, where syntactic methods can only license hopelessly weak conclusions? More specifically, what is the relationship between semantic methods of logical proof, which involves constructing models and searching for counter-examples, and standard syntactic proof theoretic methods, where a syntactic consequence relation between formulae is defined, and shown to be sound (*i.e.* not to lead from true premises to false conclusions) with respect to the semantics of the logical formulae?

The answer is disappointing: these proof methods are equivalent in the conclusions they license. Generally while insisting on the distinction between the language in which the world is described (syntax) and the described world (semantics), with respect to proof theory, logicians do not regard the syntax/semantics distinction as an appropriate dimension of difference (Scott, 1971). As we have pointed out elsewhere (Oaksford & Chater, 1993), *all* proof methods are formal and syntactic and amount to 'abortive counter-model constructions' (Hintikka, 1955, 1985). Thus, the axiomatic method, truth tables, semantic tableaux, natural deduction, and the sequent calculus are all formal proof methods which, if an argument is valid, represent abortive attempts to find a counter-model (example). Some confusion may arise, if proof theory and *model* theory are confounded, a problem we look at further below. For the moment we note that these proof methods are equivalent with respect to the inferences they are capable of making (they may, however, differ in complexity) and hence appeal to different proof procedures appears to offer no advantage to the beleaguered Logicist.

The situation is more discouraging still in the context of everyday nonmonotonic reasoning. As noted above, in deductive reasoning, showing that a conclusion follows from a set of premises involves checking

that the conclusion is true in all possible models in which the premises are true. However, in the case of non-monotonic reasoning it will be possible, by definition, for the conclusion to be false while the premises are all true. After all, in such reasoning, inferences are provisional, and conclusions may have to be retracted in the light of further information. Thus an exhaustive search for counterexamples for any non-deductive inference will inevitably be successful and no inferences will be licensed. Accordingly, it appears that, far from being readily extendible to common-sense inference, semantic methods of proof are fundamentally incompatible with it (Garnham makes just this point, in a slightly different context; see Section 6, p. 62).

It might be argued that this argument is too swift. Perhaps semantic methods of proof are applicable to non-monotonic reasoning, if there are suitable restrictions on which models are entertained (and something of this sort seems to be implicit in Garnham's discussion in Section 7). In particular, perhaps the appropriate method of proof in the nonmonotonic case is not to exhaustively search all possible models, but to entertain only the most plausible models, perhaps even just the single most plausible model. Consider, for example, the default inference from learning that Fred ate a banana to assuming that Fred peeled it first. Certainly, there are many models in which the premise is true and the conclusion false—Fred may have had the banana peeled by a friend, eaten it whole and so on. But these models are not, at least in the absence of additional information, plausible. Much more plausible is the model in which Fred peeled and ate the banana as normal. To reason successfully about these matters, it might be argued, what is required is just that a plausible, rather than an implausible model is constructed; if implausible models are constructed at all, they must be recognized as implausible and rejected.

This line of reasoning has, in Russell's phrase, all the virtues of theft over honest toil. The use of semantic methods of proof is bought at the expense of assuming as given a mechanism which can tell between plausible and implausible models—and furthermore come up with plausible models spontaneously. In other words, it presupposes a mechanism which is able to carry out inference to the best explanation—to devise and assess the plausibility of hypotheses to explain and be explained by known information. But, of course, inference to the best explanation is the very cognitive capacity for which Logicism and its allies attempt to account by adverting to methods of proof, be they syntactic or semantic. An account in which the ability to construct just the right model (the best explanation) is a primitive operation is vacuous.

Semantic methods of proof seem therefore inevitably to founder on either of these two difficulties. Without some notion of which models are plausible and which are not, it will invariably be possible to construct some (implausible) model, even for the most persuasive of common-sense inferences, and hence semantic methods will license no commonsense inferences at all. This is an even more extreme version of the problem of

weak conclusions for syntactic methods of proof: the problem of *no* conclusion. On the other hand, if some notion of plausibility of a model is presupposed, then the solution to the problem of accounting for common sense reasoning has simply been assumed rather than explained.

Garnham appears to veer towards the latter course in discussing how a model based theory of nonmonotonic reasoning might look. Rather than addressing the problem that building only a very small number of models requires some way of picking the most plausible models (that is, inferring the best explanation) Garnham argues that certain quite unexpected considerations may be sufficient to distinguish models that should and should not be considered in reasoning: '[I]t seems natural to cash this *should* in terms of what people can be expected to do, given their cognitive capacities, in particular the processing and capacity limitations of short-term working memory and the organisation and retrieval of information from long-term memory. . . . Thus, people should consider revisions of their mental models that are required by a specific piece of information that has entered working memory, from long-term memory or elsewhere' (Garnham, Section 7, p. 63). This does not, however, seem to provide any comfort for the advocate of semantic methods of proof. No doubt the organization of human memory is importantly related to human reasoning abilities; indeed, it may very well be that memory is so organized that in some way plausible models can readily be accessed, and implausible models cannot, that relevant information is fed into a short term store as required and that irrelevant information is suppressed and so on. This is just to say that human common-sense reasoning processes may be profoundly bound up with human memory, a view with which most theorists would probably concur; it goes no way at all to providing an account of how such reasoning occurs, or suggesting how such an account (presumably somehow implemented within long term memory itself) would look like a semantic method of proof.

Apart from appealing to memory, Garnham pursues a rather different line, adverting to simple strategies which can be used to guide the model building process. So, for example, '. . . revisions that falsify a conclusion consistent with the current model should not be considered, unless they are unavoidable' and 'A conclusion can be accepted (tentatively, since it is defeasible) if there is some model of the premises that will accommodate it' (Garnham, Section 7, p. 63).

Yet such proposals are entirely unable to distinguish between good and bad inferences, at least without covert assumptions concerning which models are plausible and which are not. With regard to the first principle, suppose that a reasoner who has learned that Fred ate a banana, created a model of the situation in which he peeled the banana before eating it. Suppose the reasoner then learns that Fred choked on the banana skin and had to be rushed to hospital. A natural reaction to this additional information is to overturn the tentative conclusion that Fred peeled the banana before eating it, and assume instead that he attempted to eat it

all at once. This seems more plausible than alternative models in which Fred peels and eats his banana and then eats the skin too, or whatever it might be. However, Garnham's principle does not allow such a retraction to occur, since revision of the tentative conclusion is certainly not unavoidable—just rather unlikely. Unless there is some hidden appeal to plausibility, and, we would urge, to a prior solution to the problem of inference to the best explanation, Garnham's principle will not allow us to account for the obvious common-sense conclusion.

The second principle fares no better. If any proposition which can be accommodated by some model of the premises can be accepted (albeit tentatively) then inferential anarchy appears to follow immediately. So, for example, there will be a model in which Fred eats a banana and a pig is sitting on the roof of his house (assuming no information to the contrary). Thus Garnham's second principle then licenses this (bizarre) conclusion which is (tentatively) accepted. Of course, similar reasoning can also lead to the acceptance of the opposite conclusion (although, by the first principle, the first of these to be accepted will preclude the other from being accepted). There is, of course, a very large difference between models in which there is and is not a pig on the roof—the latter will, of course, be markedly more plausible, other things being equal. But, we are arguing that plausibility is what is to be explained, and thus cannot itself be presupposed in explanation.

A natural move to dampen down the inferential chaos that Garnham's principles appear to license is to appeal to relevance—models which make specific assumptions which are entirely irrelevant to the given information (for example, models which specify the presence or absence of farmyard animals in the context of fruit eating) should be ruled out. But, appeal to relevance is just as circular as appeal to plausibility—only given the ability to successfully infer what explains what is it possible to know which facts are relevant to which other facts (see the discussion of relevance in the ***Objections and Replies*** section of the original paper).

Quite generally, the principles that Garnham invokes and others like them are inevitably doomed to fail, since they do not take into account what is being reasoned about, what it is plausible to assume, what is relevant to what, and so on; formal principles such as those we have just considered will fare no better than the rules of deductive logic in trying to account for the flexibility of common-sense inference. And of course appeals to content, plausibility or relevance are not open to the advocate of semantic methods of proof as theories of reasoning, since they assume what is to be explained.

Overall, the difference between Garnham's position and ours is that we see the problem of finding the right model as simply a restatement of the original problem of performing inference to the best explanation, whereas he treats it as a relatively straightforward matter, to be explained in terms of memory limitations, relatively simple strategies and the like. We suspect that one of the most significant contributions of recent work on knowledge

representation in artificial intelligence has been precisely that it has made clear, in painful detail, that simple formal proposals about how commonsense knowledge can be managed almost invariably rely on covert intuitions about what is and is not plausible; hence, as soon as such proposals are implemented computationally, or just formalized logically, their shortcomings become all too readily apparent.

4. *Mental Models and Mental Logics*

Our discussion of semantic methods of proof has so far been quite general, and has not been targeted at any specific proposals concerning the semantic methods of proof putatively involved in reasoning. Furthermore, we have assumed that semantic methods of proof are, like more standard syntactic methods, defined over formulae of a logical language; psychologized, this means that semantic methods of proof are defined over an internal mental logic. Semantic methods of proof are simply an alternative way of passing from premises to conclusions.

Garnham stresses that mental models theory, which he proposes as a salvation for Logicist cognitive science, is *not* a theory of mental logic, and wonders if it is this spurious identification which leads us to describe mental models theory as a semantic method of proof. Certainly, in the original paper, and in the above discussion, we have assumed that mental models theory is an alternative method of proving theorems of logic, rather than an alternative to logic itself. This is not to run together explanations of human reasoning based on mental models and those based on, say, natural deduction (*e.g.* Braine, 1978; Rips, 1983). The difference between these is precisely the difference between semantic and syntactic methods of proof (although as we mentioned above, for the logician, this is not a coherent distinction amongst proof theories). But we are assuming that both of these explanations are fundamentally explanations in terms of logical proof, though of rather different sorts. Perhaps there is no substantive disagreement here: Garnham may be using 'theory of mental logic' to apply to only syntactic proof-theoretic methods, whereas we would apply the phrase more broadly. However, it may be that the importance which Garnham attaches to this objection stems from the view that mental models theory should not be assimilated with proof-theoretic methods since it is very different in character, in ways that we have failed to appreciate. For example, he notes that our discussion 'equivocates on the term "logic". Much of the time they write as if the only hypothesis worthy of consideration is that the system of operations underlying human reasoning corresponds to some established logical system (*e.g.* one of the standard nonmonotonic logics) . . . [yet] there are many logics that cannot be reduced to first-order logic . . . and which can be formalised model-theoretically. And although extended model theory has its primary applications in mathematics, there are certainly aspects of everyday reasoning . . . that

call for formalisations which are model-theoretic and not proof-theoretic in nature.' (Garnham, p. 63, fn. 14).

The thrust of this disagreement is perhaps not entirely clear. Initially, we are held to equivocate on the term 'logic'; yet the follow up point is that certain logics, which may be important for understanding everyday reasoning cannot be formalized proof theoretically. So it seems that the term 'logic' is not in dispute after all; Garnham has just as wide a notion of logic in mind as we do. Presumably this means that the question of whether or not a model based account is a theory of mental logic is similarly a red herring. The substantial claim appears to be that model-based accounts of reasoning are, in principle, more powerful than proof theoretic methods.

As discussed above, this claim is not correct since the distinction between semantic and syntactic methods of proof is not one that can generally be enforced. As we also mentioned above, the reason that the opposite view can seem plausible is due to a conflation between model theoretic semantics (which provides *truth conditions* for formulae of a logical language) and mental models theory (which provides an inference mechanism). Providing a semantics and providing an inference mechanism are, of course, very different things (see *e.g.* Hintikka, 1985)—yet in Garnham's discussion the term model-theoretic is used to apply to both. When Garnham notes that many logics can only be formalized model theoretically, what is meant is that while higher order logics can be given a semantics in terms of abstract, set-theoretic structures, a syntactic proof theory which captures all and only the valid inferences licensed by that semantics cannot be provided. The standard semantic notion of validity, that all models in which the premises are true must also make the conclusion true, can be applied using such model structures, but the class of semantically valid inferences cannot be captured using proof theoretic rules—there will, in particular, be semantically valid inferences which any proof theory will be unable to capture. Thus, it will not be possible to construct a mechanized proof theory which will capture all and only semantically valid inferences.

This by no means implies that mental models can fair any better, however. Indeed, for incomplete logics there is provably *no* mechanism, based on whatever principles, which will capture all and only valid inferences (Boolos & Jeffrey, 1980). In practice, semantic methods of proof become entirely unworkable as the logic becomes more complex, since the space of possible models becomes enormously large (for example, in second order logic, involving each possible *set* of objects corresponding to a predicate; in modal logics, involving the interpretation of a term across *each possible world* may have to be considered). Thus practical attempts to build reasoning systems using higher order logics have generally attempted to implement incomplete syntactic proof theories rather than search for counterexamples through gigantic sets of possible models. In particular, this means that the mechanisms of mental models theory appear,

in general, *less* well suited than traditional syntactic proof theory to dealing with the kind of reasoning that Garnham notes is important in formalizing everyday reasoning.

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

The introduction of arbitrary exemplars highlights the lack of an appropriate meta-theory for mental models (Oaksford & Chater, 1993). Mental models theorists provide no exposition of the rules which guarantee that no illegitimate assumptions are introduced in a proof. This does not mean that any particular derivation using mental models has made such assumptions. Nonetheless, guaranteeing the validity of an argument depends on ensuring that in a particular derivation such an assumption *could* not be made. Hence explicit procedures to prevent this happening need to be provided. In their absence there is no guarantee (*i.e.* no proof) that the procedures for manipulating mental models preserve validity. That is, it is not known whether, relative to the standard interpretation of predicate logic, mental models provides a *sound* logical system.

While soundness is unresolved, there are strong reasons to suppose that mental models theory is not *complete* with respect to standard logic, *i.e.* while all inferences licensed by mental models may be licensed by standard logic (soundness) the converse is not the case. Other *graphical* methods of proof, such as Venn diagrams or Euler's circles, are restricted in their *expressiveness* due to physical limitations on the notation. Venn diagrams for example, can only be used to represent arguments employing 4 or less *monadic* predicates, *i.e.* predicates of only one variable (Quine, 1959). They therefore only capture a small subset of logic. While mental models have been used to represent relations, *i.e.* predicates of more than one variable, there is no reason to suppose that mental models will not be subject to analogous limitations.

The employment of arbitrary exemplars is also central to providing a tractable model based proof procedure (see Oaksford & Chater, 1993). However, in the absence of complexity results for the algorithms which manipulate mental models, a demonstration that mental models can avoid the intractability which bedevils the syntactic approach to nonmonotonic reasoning remains wanting.

It is perhaps because of a conflation between set-theoretic and mental models, that mental models accounts do not generally attempt to define a

semantics for their mental models notation. For example, the following, from the most recent text that Garnham cites (Johnson-Laird & Byrne, 1991), are the mental model representations of three possible interpretations of the conditional sentences employed in Wason's (1966) selection task.

```
[A]  2      [A]  [2]      [A]  2
 ...         ...               not-2
```

This is a complex notation, the precise meaning of which is only specified intuitively. Yet the notation of mental models theory stands as much in need of semantics as the notation of standard logic. Without a well-defined semantics it is impossible to know whether or not rules postulated for manipulating such models are valid. In this sense, then it could perhaps be said that mental models theory, in its current incarnation, can be distinguished from logic, in being less fully formalized. It seems unlikely however that this distinction is one which mental models theory will find to its advantage.

5. Conclusions

In our original paper, we argued that a Logicist cognitive science of central processes cannot account for the common-sense inferences that people draw, and cannot be tractably implemented. We argued furthermore that positions closely related to Logicism, including those, such as mental models theory, which use semantic rather than syntactic methods of proof, equally succumb to these problems. We have found no persuasive reason to alter this conclusion in the light of Garnham's discussion.

Finally, while we agree with Garnham that the question of whether connectionist systems can help in providing tractable theories of everyday reasoning is undecided, we do not share his pessimism concerning the final answer. The general principles of how everyday reasoning may be accommodated within neural networks are likely to involve the exploitation of a more complex dynamics and there is much recent progress in this area relevant to providing tractable accounts of real human inference (see, Shastri & Ajjanagadde (in press); and the papers in Oaksford & Brown (in press)).

Department of Psychology
University of Edinburgh
7 George Square
Edinburgh EH8 9LZ
UK

Department of Psychology
University of Wales, Bangor
Gwynedd LL57 2DG
UK

References

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.
Fodor, J.A. 1975: *The Language of Thought*. New York: Thomas Crowell.
Fodor, J.A. 1983: *The Modularity of Mind*. Cambridge, MA.: MIT Press.
Fodor, J.A. and Pylyshyn, Z.W. 1988: Connectionism and Cognitive Architecture: A Critical Analysis. *Cognition*, 28, 3–71.
Ginsberg, M.L. 1987: *Readings in Nonmonotonic Reasoning*. Los Altos, CA.: Morgan Kaufman.
Goodman, N. 1983: *Fact, Fiction and Forecast*. 4th Edition, Cambridge, MA.: Harvard University Press. (Originally published 1954.)
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.
Hintikka, J. 1955: Form and Content in Quantification Theory. *Acta Philosophica Fennica*, 8, 11–55.
Hintikka, J. 1985: Mental Methods, Semantical Games, and Varieties of Intelligence. Unpublished Manuscript, University of Florida.
Holland, J.H., Holyoak, K.J., Nisbett, R.E. and Thagard, P. 1986: *Induction: Processes of Inference, Learning and Discovery*. Cambridge, MA.: MIT Press.
Johnson-Laird, P.N. 1983: *Mental Models*. Cambridge: Cambridge University Press.
Johnson-Laird, P.N. and Byrne, R.M.J. 1991: *Deduction*. Hillsdale, N.J.: Lawrence Erlbaum Associates.
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 & Co.
Oaksford, M. and Brown, G.D.A. In Press. *Neurodynamics and Psychology*. London: Academic Press.
Oaksford, M. and Chater, N. 1991: Against Logicist Cognitive Science. *Mind and Language*, 6, 1–38.
Oaksford, M. and Chater, N. 1992a: Bounded Rationality in Taking Risks and Drawing Inferences. *Theory and Psychology*, 2, 225–30.
Oaksford, M. and Chater, N. 1992b: Reasoning Theories and Bounded Rationality. In K. Manktelow and D. Over (eds.), *Rationality*. London: Routledge.
Oaksford, M., Chater, N. and Stenning, K. 1990: Connectionism, Classical Cognitive Science and Experimental Psychology. *AI and Society*, 4, 73–90 (Also in A. Clark and R. Lutz (eds.), *Connectionism in Context*. Berlin: Springer-Verlag, 1992, 57–74.)
Oaksford, M. and Stenning, K. 1992: Reasoning With Conditionals Containing Negated Constituents. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 18, 835–54.
Pylyshyn, Z.W. 1984: *Computation and Cognition: Toward a Foundation for Cognitive Science*. Montgomery, VT.: Bradford Books.

Quine, W.V.O. 1959: *Methods of Logic*. New York: Holt, Rinehart and Winston.
Rips, L.J. 1983: Cognitive Processes in Propositional Reasoning. *Psychological Review*, 90, 38–71.
Scott, D. 1971: On Engendering an Illusion of Understanding. *Journal of Philosophy*, 68, 787–807.
Shastri, L. and Ajjanagadde, V. In Press: From Simple Associations to Systematic Reasoning: A Connectionist Representation of Rules, Variables, and Dynamics Bindings Using Temporal Synchrony. *Behavioural and Brain Sciences*.
Shoam, Y. 1987: *Reasoning About Change*. Cambridge, MA.: MIT Press.
Shoam, Y. 1988: Efficient Reasoning about Rich Temporal Domains. *Journal of Philosophical Logic*, 17, 443–74.
Thagard, P. 1988: *Computational Philosophy of Science*. Cambridge, MA.: MIT Press.
Wason, P.C. 1966: Reasoning. In B. Foss (ed.), *New Horizons in Psychology*. Harmondsworth, Middlesex: Penguin.