In W. Ramsey, S. Stich and D. Rumelhart (eds), *Philosophy and Connectionist Theory* (Hillsdale, NJ: Lawrence Erlbaum Associates), 229–257.

Concepts, Connectionism, and the Language of Thought

Introduction

The aim of this paper is to demonstrate a *prima facie* tension between our commonsense conception of ourselves as thinkers and the connectionist programme for modelling cognitive processes. The language of thought hypothesis plays a pivotal role. The connectionist paradigm is opposed to the language of thought; and there is an argument for the language of thought that draws on features of the commonsense scheme of thoughts, concepts, and inference. Most of the paper (Sections 3-7) is taken up with the argument for the language of thought hypothesis. The argument for an opposition between connectionism and the language of thought comes towards the end (Section 8), along with some discussion of the potential eliminativist consequences (Sections 9 and 10).

Jerry Fodor has been bombarding us with arguments for the language of thought (LOT) hypothesis, from his book *The Language of Thought* (1975), through to *Psychosemantics* (1987), and Fodor and Pylyshyn's (1988) attack on connectionism and its followers. The argument to be presented here has close affinities to some of Fodor's recent arguments, but its character is more *a prioristic*.

Some philosophers may resist all these arguments, because they have strong intuitive reservations about the very idea of a LOT. Our first task (Sections 1 and 2) is to consider some of those reservations.

1. Order out of chaos

In *Zettel*, Wittgenstein writes:

> 608. No supposition seems to me more natural than that there is no process in the brain correlated with associating or with thinking; so that it would be imposible to read off thought-processes from brain-processes. I mean this: if I talk or write there is, I assume, a system of impulses going out from my brain and correlated with my spoken or written thoughts. But why should the *system* continue further in the direction of the centre? Why should this order not proceed, so to speak, out of chaos? . . .
>
> 609. It is thus perfectly possible that certain psychological phenomena *cannot* be investigated physiologically, because physiologically nothing corresponds to them.

What this suggests to some philosophers is that they have no business prejudging the nature of the causes of behaviour.

Tracts of behaviour exhibit enough 'system' to support radical interpretation (Davidson 1973) - to permit the adoption of the intentional stance (Dennett 1971, 1981). In accordance with the interpretive strategy, we cast a net of psychological description over the writhing mass of behaviour of an intentional system, such as one of us. The psychological description has a structure, namely, the structure of the public language sentences that are used following the psychological verbs: 'she believes that . . . ', 'she desires that . . . ', 'she intends that . . . '. Doubtless, the tracts of behaviour that are so interpreted have causal explanations. But, we should not assume that the causes of behaviour - whether they be characterised physiologically, or in any other way - must have an articulation matching the structure of that psychological description. This, in

contemporary dress, is the lesson that some philosophers are inclined to draw from *Zettel* 608-9.

It is little wonder, then, that philosophers of a Wittgensteinian persuasion should sometimes feel drawn to connectionism. For a connectionist network whose performance of a cognitive task is mediated by connections to and from a mass of individually uninterpretable hidden units seems perfectly to capture the idea of order proceeding, so to speak, out of chaos. In contrast, the LOT hypothesis seems, from this perspective, like a quite gratuitious assumption that the 'system' must continue 'further in the direction of the centre'.

However, this first kind of reservation about the LOT hypothesis is evidently open to argument. It may be that good arguments can be mounted for the view that it is severely improbable that richly structured behaviour should be forthcoming without causal antecedents that exhibit the structure of a LOT. The constraint under which such arguments operate is simply that they should address, and not ignore, the possibility of alternative patterns of causal antecedents that is urged by connectionism.

There is a second kind of reservation that is potentially more serious. It is that the LOT hypothesis involves a regress of one kind or another.

A sentence is *inter alia* a syntactic object. When a sentence of a public language is presented to someone who understands the language, he is able to assign a meaning to the sentence, and thus to take the sentence as a semantic object. So, if thinking involves tokens of sentences in the LOT, for whom are these sentences objects? To whom are these inner sentences presented *inter alia* as syntactic items standing in need of interpretation?

To answer that the LOT sentences are presented in this way to the thinker himself seems quite unsatisfactory, for it involves a regress of languages: a metalanguage of thought in which to think about the sentences in one's language of thought, a metametalanguage, and so on. On the other hand, the answer that the sentences of a thinker's LOT are presented as syntactic items to a little man who reads and understands what is written on some inner blackboard also seems hopelessly regressive. Since understanding itself involves thinking, the little man would need to have his own inner blackboard, and a further, smaller, little man to read what is inscribed upon it.

Once this worry takes hold, it can come to seem that the thesis that thinking requires an inner language of thought is about as philosophically disreputable as the thesis that picking a red flower requires an inner collection of colour cards (Wittgenstein, *Blue Book* (1969), p. 3).

2. Avoiding the threat of regress

But the worry should not be allowed to take hold. The LOT hypothesis is not regressive.

In order to see this, we need to know what a minimal version of the hypothesis says. The statement that thoughts have syntactic properties is, by itself, unclear. We must distinguish between two senses of the word 'thought', and say which properties count as syntactic.

The word 'thought' can be used to apply to thought contents or to thought states. If thought contents are the focus of attention, then it is easy to trivialise the LOT hypothesis. For someone might hold that thought contents are canonically specified in a public language like English, and that they simply inherit a syntactic articulation from

their linguistic specification. But the LOT hypothesis is not a claim about the structure of our public language descriptions of thinkers. It concerns, not our attributions of mental states, but the structure of the states thus attributed.

If we were primarily concerned with thought attributions in a public language, then the idea of syntactic properties would be unproblematic. But what does it mean to say that certain states of a thinking person have syntactic properties?

From Fodor's work (1987a, pp. 16-21), we can extract three conditions upon syntactic properties. First, a syntactic property is a (higher order) physical property. Second, syntax is systematically related to semantics. Third, a syntactic property is a determinant of causal role or causal powers.

It would be entirely fair to complain that these three conditions are not utterly transparent. For example, Fodor says that shape is the right sort of property to be a syntactic property; and of course, shape is an intrinsic property. So, we might ask whether the first condition is intended to require that syntactic properties of a state are intrinsic, rather than relational; whether this requirement is, rather, an intended consequence of the third condition; or whether the definition of syntax is not itself intended to rule out relational properties. (For a discussion of the notions of syntactic and formal properties, see Devitt 1989.) Likewise, we might ask whether causal powers include both active and passive powers - whether causal role includes both upstream and downstream role.

Let us settle this latter question in favour of just active powers, or downstream causal role, and - ignoring other unclarities - content ourselves with a minimal notion of a syntactic property as a *physical* property that is both *systematically related to semantics* and a *determinant of causal consequences*.

The argument for the LOT to be presented here concerns certain states of thinkers. They are states that have semantic properties, and are the inputs to various processors or mechanisms. The conclusion of the argument is that these states have properties that are correlated with their semantic properties and engage those mechanisms. According to the minimal notion of syntax, these properties count as syntactic.

It is easy to see that there is nothing potentially regressive in that conclusion. For the states that have syntactic properties are not presented to anyone - not to the thinker and not to a little inner man - as standing in need of interpretation. The processors which the states engage are not in the position of someone who is presented with a syntactic item but who does not know what to do with it until he knows what it means.

On this point Fodor (1987b, p. 67) says :

> [the formulas of LOT] - unlike those of German - can have their effects on the course of thoughts without having to be understood. This is precisely because - according to the computational story - the psychological effects of Mentalese [LOT] formulas are mediated by their syntactic/intrinsic properties (rather than their semantic/relational properties). This is the trick that computational psychologists use to get the goodness out of postulating a language of thought without having the traditional problem of a regress of languages and interpreters.

It is possible that someone might complain that, precisely because the LOT hypothesis does not give rise to a regress, the terms 'syntax' and 'language' are misleading since they suggest the case of public language.

There is, perhaps, a hint of this complaint in Barwise's response (1987, p. 83):

> I now realise that for Fodor, the features of language that make 'language of thought' an appropriate metaphor have to do with combinatorial-structural properties, whereas what Perry and I have reacted to in the use of this metaphor is the idea that these 'expressions' are things that have to be 'read' or otherwise 'understood' the way expressions of a language used for communication between agents have to be in order to have significance.

To the extent that this complaint is justified, the conclusion of the argument may not strictly deserve to be called the truth of the *language* of thought hypothesis. But, given that we are explicit about which properties we count as syntactic, and given that it is agreed that the minimal LOT hypothesis is not regressive, there is little here to fight over. Nor is there any need for terminological confusion, provided that we distinguish the LOT from public languages.

With all reservations about regress now set aside, we can turn to the structure of the argument itself. It comes in two main stages. The first stage (Sections 3 and 4) is an argument for a conditional claim: If a cognitive process is systematic - in a sense to be defined - then the inputs to that process have syntactic structure - in the minimal sense just characterised. The second stage (Sections 5 and 6) is an argument for the claim that being a thinker - a believer, a deployer of concepts - involves the systematicity of certain inferential transitions amongst thoughts. Consequently, the inputs to those transitions - thoughts - have syntactic structure; that is, there is a language of thought. (Section 7 deals with two objections to the second stage.)

3. Systematic cognitive processes

The first stage of the argument involves a conditional claim: If a cognitive process is systematic, then the inputs to that process have syntactic structure. Some of the terms in this claim need a little explanation.

For present purposes, a *cognitive* process is one whose input states have a semantic description, and whose output states have either a semantic description or an action description. A cognitive process thus moves from information to information, or from information to action.

The idea of a *systematic* process that is used in the argument is an essentially relative one: a process is *systematic relative to* a pattern in its input-output relation. Suppose that a generalisation G describes a pattern to be found in the input-output relation of some physical system. If we consider several input-output pairs that exhibit the common pattern, then we can ask whether the several input-output transitions have a common causal explanation corresponding to the common pattern that they instantiate. If there is a common causal explanation, then we can say that the process leading from those input states to output states is causally systematic relative to the pattern described by G.

The drinks machine

Consider the following very simple example. There is a machine that produces coffee or tea with or without milk; the output states of the machine are states of delivering drinks of one of four kinds. The input states of the machine are states of having a token of one of four kinds in its slot.

The four kinds of token are these. There are square red tokens, square blue tokens, round red tokens, and round blue tokens. If a square red token is in the slot, then the

machine delivers coffee with milk. If there is a square blue token in the slot, then the machine delivers coffee without milk. If there is a round red token in the slot, then tea with milk is delivered. And if there is a round blue token in the slot, then tea without milk is the result.

Under these descriptions of input and output states, there is a clear pattern to be discerned in the input-output relation for the drinks machine. Whatever the colour of the token (whether it is red or blue), if it is square then coffee is delivered, whereas if it is round then tea is delivered. And whatever the shape of the token (whether it is square or round), if it is red then a drink with milk is delivered, while if it is blue then a drink without milk is delivered.

We can ask whether the process that mediates between input and output states is causally systematic relative to the pattern described by each of the four little generalisations about the machine's input-output relation.

Thus, for example, is there a common explanation for the delivery of coffee consequent upon the input state of having a red square in the slot and the input state of having a blue square in the slot? Is there a common processor or mechanism responsible for mediating these two transitions? Likewise, is there a common explanation for the inclusion of milk in the drink delivered when a square red token or a round red token is inserted in the slot?

The answers to these questions are not determined by the facts about the input-output relation, but by facts about the internal architecture of the drinks machine. One possible configuration would have, within the machine, four autonomous and totally dedicated drink producing devices: one activated by each of the four possible input states. Another possible internal configuration would have three component devices. First, there would be a device that is activated by either a square red token or a square blue token in the slot, and produces coffee. A second device would produce tea if there is a round red or round blue token in the slot. And a third device would add milk to the drink produced if there is a round or square red token in the slot but refrain from adding milk if there is a round or square blue token in the slot.

These two configurations would license different answers to the question about causal systematicity. The operation of a drinks machine with the first configuration is not causally systematic relative to the input-output patterns that we discerned, whereas the operation of a machine with the second configuration is systematic.

As this example illustrates, if we think of a physical system as containing various subsystems or mechanisms, then the requirement for causal systematicity relative to the pattern described by G is that there should be a mechanism whose presence in the system explains all the input-output transitions that conform to the pattern described by G. It is not sufficient that this common mechanism should merely figure as a component somewhere along the way in the several transitions. Rather, the common mechanism should actually mediate between inputs and outputs in accordance with G.

The example also illustrates that conformity to an input-output pattern is in no way sufficient for causal systematicity, as that notion is deployed in the argument for the LOT. There can be two systems with the same input-output relation, where the processing in one system is causally systematic relative to some pattern in that input-output relation, while the processing in the other system is not systematic relative to that

pattern. The distinction here is one of which we make widespread use in our descriptions of complex systems.

The conditional claim to be established in this first stage of the argument concerns systematic cognitive processes. These are processes that are systematic relative to patterns revealed under semantic descriptions of the input and output states (or semantic descriptions of the input states and action descriptions of the output states). The semantic description of the input states is crucial to the conditional claim since, according to the notion of syntax that we are using, there are no syntactic properties without semantic properties. (For other purposes, a different notion of syntax would be appropriate. See, for example, Stich 1983; and, for some problems over the idea of syntax without semantics, Crane 1990.)

For a simple example in which the input states do have semantic descriptions, we can return to the drinks machine.

The drinks machine again

Let us speak in a pretheoretical way, and say that the machine's input state of having a square red token in the slot means that the client wants coffee with milk, the presence of a square blue token in the slot means that the client wants coffee without milk, and so on.

Given these semantic descriptions of the input states and the action descriptions of the output states, we can redescribe the pattern in the input-output relation of the drinks machine. If the input state means that the client wants coffee (whether with or without milk), then the output state is a delivery of coffee. Similarly, if the input state means that the client wants a drink with milk (whether it be tea or coffee), then milk is included in the drink that is delivered. And so on.

As before, we can ask whether the operation of the machine is causally systematic relative to each of the patterns described by these generalisations. One possible internal configuration of the machine will warrant a negative answer; another will warrant an affirmative answer.

The sentence interpreter

Consider now the cognitive process of understanding some English sentences. To be more accurate, what is to be considered is the process that begins with a state registering the information that a particular sentence has been uttered, and ends with a state registering information as to what has been said - what message has been conveyed.

If you understand the three sentences, 'Martin is tired', 'Martin is tall', 'Martin is drunk', then in each case you end up knowing that what has been said concerns this person here. We can describe a pattern in the input-output relation: If the input state registers the utterance of a sentence that contains the name 'Martin', then the output state means that what was said was about this person here. Likewise, we can describe patterns relating to the messages conveyed by other sentences: 'Martin is tired', 'Andy is tired', 'Frank is tired'. If an input state registers the utterance of any of these sentences, then the output state that is produced means that what was said was to the effect that someone (whether it be this person here, or . . .) is tired.

The causal systematicity of this cognitive process requires more than just conformity to these patterns. Systematicity relative to these generalisations requires that, corresponding to each pattern (in extension) there should be a common mechanism whose

presence explains the aspect of input-output transitions that is captured in that pattern. Within the physical system that performs the transitions that we are calling sentence interpretation, there should be a component mechanism that is responsible for mediating transitions between input states registering the uttererance of sentences containing the name 'Martin' and output states that concern this person here. Likewise, there should be a component mechanism that is responsible for mediating the several transitions from input states that concern sentences containing the predicate 'is tired'. The interpretation of the sentence 'Martin is tired' will then be the joint product of those two mechanisms. Causal systematicity thus requires real commonality of process.

Knowledge of rules

The idea of causal systematicity is also involved in the account that I would give of knowledge of rules (Davies 1987, 1989, 1990b, 1990c). Where there is causal systematicity relative to a pattern revealed under a semantic description of the input and output states, there the system has knowledge of the rule or generalisation describing that pattern.

The example of the sentence interpreter provides a straightforward case. Knowledge of the rule that sentences containing the name 'Martin' convey propositions about this person here does not require the ability to formulate explicitly the thought that that is indeed a rule of the language in question. What it does require is causal systematicity relative to the input-output pattern described by that rule.

The conditional claim with which the first stage of the argument is concerned can then be stated in terms of knowledge of rules: If a cognitive processing system embodies knowledge of a rule, then the input states of the system have syntactic structure. This conditional claim does not say that if a process is causally systematic, and so involves knowledge of a rule, then the process operates in virtue of an explicit syntactic encoding of the rule that is known. For all that the conditional claim says (and for all that Fodor says - see 1985, p. 95, 1987a, p. 25) the standing condition of knowledge of a rule can be realised just as well by the presence of a component processor as by the presence of an explicit representation. The consequent of the conditional claim concerns just the input states of cognitive systems.

4. From system to syntax

We now have some grasp upon both the antecedent and the consequent of the conditional claim. The notion of causal systematicity is a relative one, and the cases that concern us involve systematicity relative to patterns that are revealed when the input states (at least) are given semantic descriptions. The minimal notion of a syntactic property that we are using is also a relative one; in fact, it is doubly relative.

First, what counts as a syntactic property depends upon what semantic properties are present, since a syntactic property must be systematically related to semantics. Second, what counts as a syntactic property of an input state depends upon the actual constitution of the machine of which it is an input state. For a syntactic property must be a determinant of causal powers, and a property to which the operation of one machine is sensitive may be quite irrelevant to the operation of another machine.

Of course, it is one thing to understand a claim, and quite another to have an argument for its truth. We can proceed a good part of the way towards seeing why the conditional

claim is true if we return to the first example of the drinks machine, in which the input states are described simply in terms of the shape and colour of the tokens in the slot. For we can observe that causal systematicity of process imposes requirements upon the causal properties of the input states.

Suppose that the operation of the drinks machine is causally systematic. Then there is *inter alia*, as a component of the drinks machine, a common mechanism that operates to mediate the transition from either a square red token or a square blue token in the slot to the delivery of coffee. But then there must be some property shared by - and distinctive of - those input states which is causally adequate to engage that mechanism.

There must also be a causally relevant difference between those two input states, since one state engages the milk introducing mechanism while the other does not. Indeed, in order to engage that mechanism, the input state of having a square red token in the slot must have some causal property in common with the input state of having a round red token in the slot - a property not shared by the other two input states. In short, the input states exhibit patterns of recurrent properties that are determinants of the causal consequences of those states in the context of the drinks machine.

It is, of course, an empirical question what the causally salient properties of the input states are. All that causal systematicity requires is that the operative properties of the input states should correlate with the properties cited in the descriptions of the input-output patterns. It might be that the squareness of tokens in the slot is what engages the coffee mechanism, and that the redness of tokens is what engages the milk mechanism. But it might also be that the square tokens or the red tokens have something else in common, such as a distinctive mass, a distinctive chemical composition, or a distinctive inscription upon them.

The properties required by causal systematicity in the first example of the drinks machine do not yet qualify as syntactic properties, since no semantic properties have been introduced for them to be correlated with. But we can take the final step towards seeing why the conditional claim is true if we now consider the second example involving the drinks machine. In that example, the input-output pattern is revealed under a semantic description of the input states.

So, suppose that the operation of the drinks machine is causally systematic relative to the patterns that are revealed under semantic descriptions of the input states. This is to say that the machine has knowledge of rules such as: Deliver coffee, given that the client wants coffee.

This causal systematicity imposes a requirement that the input states that mean that the client wants coffee (whether with or without milk) should have a causal property in common, in virtue of which those input states engage the coffee producing mechanism. Similarly, the input states that mean that the client wants a drink with milk should have in common a property that engages the milk introducing mechanism. There is no particular requirement as to what these properties should be; they might or might not be the squareness and redness, respectively, of the token in the slot. But, over the range of input states, they have to correlate with meaning that the client wants coffee and meaning that the client wants a drink with milk respectively. Exactly similar considerations apply to the example of the sentence interpreter.

In short, what we see is that causal systematicity relative to semantic input-output patterns (or equivalently, knowledge of rules) requires that the input states of the

machine should have properties that are correlated with their semantic properties, and are determinants of the causal consequences of those states given the internal constitution of the machine. Since these will surely be physical properties, they will meet all three conditions on syntactic properties.

There are two points to notice about the conditional claim. The first point is that the complexity of the syntax of input states may be very modest indeed. For example, in the case of the drinks machine, the formal language of its input states has just four primitive symbols and one binary operation; and the operation does not even distinguish the order of constituents. The second point is that not every aspect of semantic content is required to be articulated syntactically (*cf.* Perry 1986). The argument for the conditional claim requires no syntactic property corresponding to an aspect of semantic content that is constant over all input states of the machine (such as that all input states mean that the client wants something). The drinks machine is dedicated to the wants of the client; and task dedication permits syntactic inarticulateness.

So much, then, for the first stage of the argument for the LOT hypothesis. Its plausibility depends upon two things. A minimal notion of syntactic property is used in the consequent of the conditional; and the notion of causal systematicity that is used in the antecedent requires much more than just that a pattern (in extension) should be exhibited by an input-output relation. In short, the truth of the conditional claim is secured by having a relatively strong antecedent and a relatively weak consequent. The price of this strategy is, of course, that it increases the burden upon the second stage of the argument.

The task of the second stage of the argument is to uncover, in the commonsense scheme of thought, concepts, and inference, a commitment to causal systematicity of cognitive processes. That second stage depends upon a neo-Fregean conception of thoughts.

5. The structure of thought

Thoughts are states with semantic content, and consequently with truth conditions. But thoughts are not the only semantically evaluable states in the world; nor are they the only psychological states that have semantic content. Having content is a feature that is shared by thoughts, by certain patterns of sound waves and marks on paper, by states of the visual system of humans and other animals, by patterns of tree rings, and by states of room thermostats.

This is not to say that thoughts have content in just the same way as these other states. Someone might hope that there can be a unified theory of all these contents, but there are reasons to think that we need to distinguish between information content - for which some causal-cum-telelogical theory might be right - and mental content (the content of propositional attitudes) - for which such a theory is inadequate.

If this is right, then, amongst the psychological states of a person, we need to distinguish between states with mental content and states with mere informational content - between propositional attitude states and *subdoxastic* states (Stich 1978). And a substantial project in the philosophy of psychology is to give a principled account of this distinction.

One way to commence on that project is to focus upon the fact that thoughts - and attitude states in general - are states whose semantic content is *conceptualised* content. A

person who is in such a state *ipso facto* deploys the constituent concepts of the content of that state (*cf.* Davies 1989). This is not so for states of early visual processing, for example.

The 'ipso facto' is important; for certainly a person may have a thought about the content of a state of visual processing, and so conceptualise the content of that state. Similarly, a theoretical linguist may have a thought about the information content of some state of the language system. But having those thoughts is not essential to being in the respective states.

The content of thoughts is conceptualised content. To entertain a thought, to hold a belief, or to frame an hypothesis, involves deployment of concepts. Thus, no one can entertain a thought with a particular content without grasping the constituent concepts of that content. Furthermore, for a thinker to have the concept of being F, the thinker must *know what it is for* an object to be F - that is, know what it is for an arbitrary object to be F. (This epistemic requirement for possession of the concept of being F has an analogue for thoughts about particular objects; namely, that the thinker should *know which* object is in question. Gareth Evans (1982, p. 65) calls this *Russell's Principle*. As Evans points out (*ibid*, pp. 76-9), these requirements involve rejection of the *Photograph Model* of mental representation.)

Putting these ideas together, we arrive at an important neo-Fregean consequence. To entertain the thought that object a is F a thinker must have the concept of being F. If a thinker has that concept, then the thinker knows what it is for an arbitrary object to be F. So, if a thinker thinks that a is F and is able to think about the object b, then the thinker is able to entertain the thought - to frame the hypothesis - that b is F.

This consequence is, in effect, what Evans (*ibid*, p. 104) calls the *Generality Constraint*; and it has as an immediate consequence (perhaps not properly distinguishable from the Generality Constraint itself) a *closure condition* on the domain of thought contents available to a thinker.

If a thinker can be credited with the thought that object a is F and the thought that object b is G, then that thinker has the conceptual resources for also entertaining the thought that a is G and the thought that b is F. Similarly, if a thinker can be credited with the thought that a is R to b, then that thinker has the conceptual resources for also entertaining the thought that b is R to a. The domain of thought contents available to a thinker is closed under recombination of conceptual constituents.

Thoughts are states with semantic content, and these contents are of a special kind that is subject to the Generality Constraint and thus to the closure condition. These are two important neo-Fregean claims. But they leave us some way from causally systematic processes; and they do not themselves permit a direct argument for the LOT.

Semantic content and the closure condition

Consider first the claim that thoughts are states with semantic content. Certainly there is no argument from the mere fact of semantic content to the LOT hypothesis.

Suppose that a creature evolves in an environment where the prime predatory danger typically arises after a hawk dives on a beetle. Suppose that the creature develops a detector for just this scenario: a hawk-diving-on-beetle detector. Suppose that the detector operates by being sensitive to overall aspects of the threatening scenario; and not by being composed *inter alia* from a hawk detector and a beetle detector. It is immensely

plausible that, for a causal-cum-teleological notion of information content, there is a state of this creature with the semantic content that a hawk is diving on a beetle. But there is no reason at all to suppose that the state has a syntactic constituent structure.

Consider second the claim that thought contents are subject to the closure condition. There is no utterly compelling argument from the closure condition on semantic contents to the LOT hypothesis.

One manifestation of the closure property is that if a system has a state with the content that a is R to b, then it also has a state with the content that b is R to a. But the availability of states with these contents does not require a syntactically structured vehicle for semantic content.

Thus suppose that, for whatever reason, our creature with a hawk-diving-on-beetle detector also develops a second detector for a second threatening scenario. Suppose that danger is often just around the corner when a beetle dives on a hawk; and that the creature consequently develops a beetle-diving-on-hawk detector. Let it be that these two proprietary detectors - with their downstream processors that produce appropriate evasive behaviour - are causally autonomous from each other; indeed, we could think of each one as a module within the creature's total information processing system.

This example has been set up so that there is no common syntactic constituent in these two information registering states. There is no syntactic symbol meaning *hawk* that is implicated in the two states, for example. For the two states to have a syntactic constituent in common would require there to be some common property of the two states which is systematically related both to the semantic content of the two states and to the causal consequences of the two states. There is, to be sure, a common property of the two states which is related to their semantic contents; namely, a complex relational property having to do with the causal antecedents of the states. But this property is not directly implicated in the production of the causal consequences of the states.

This example of the two detectors is, of course, a mere toy. The idea behind the example can be extended to some other, more complex, toys, such as the sensori-motor coordination system in Paul Churchland's (1986) crab. Also, in the context of connectionist representation, the idea can be applied to the binding units employed in simple tensor product schemes (Hinton, McClelland and Rumelhart 1986; Smolensky 1987). But, with or without further examples, the principle is clear. *De facto* compliance with the closure condition does not inevitably require syntax.

It might be responded here that it is significant that all the examples that are supposed to illustrate this principle share the feature of being only toy examples. The claim might be made that once we attempt to meet the closure condition for a suitably rich set of semantic contents, without overreaching the available computational resources, we shall inevitably have to make use of some syntactic articulation. For, it might be asked, *how else* could we do it?

This is quite a potent challenge. We are certainly not obliged to deny that - properly developed - it can establish a strong plausibility consideration in favour of the LOT hypothesis. Nevertheless, a 'How else?' challenge is always open to the risk that someone will respond: 'Like so'. This is what connectionists, for example, do.

The LOT hypothesis is supposed to play a pivotal role in an argument for a *prima facie* tension between our conception of ourselves as thinkers and the connectionist programme. Consequently, the argument from our commonsense conception to the LOT

must not appear to beg the question against connectionism. What is needed, and what is being offered here, is an argument that is more direct and *a prioristic* than any 'How else?' challenge.

To the extent that we do not rest content with a 'How else?' challenge, we shall also not rely upon an argument by analogy that proceeds as follows.

The semantic contents of sentences of a natural language meet a closure condition. If, for example, there is a sentence that means that a is R to b, then there is also a sentence that means that b is R to a. The sentences of a natural language meet this closure condition by having a syntactic constituent structure. The contents of thoughts meet a similar closure condition. Therefore, by analogy, thoughts have syntactic constituent structure, too.

The analogy between the meanings of natural language sentences and the contents of thoughts is not perfect. After all, it is possible to have mere phrasebook mastery of a (fragment of) a language; whereas it is not possible to have phrasebook mastery of thoughts (*cf.* Evans 1982, p. 102). But, once again, it is not necessary for us to hold that the argument is totally without merit. It is simply that the argument by analogy cannot serve our dialectical purpose; the most that it can achieve is to establish a plausibility consideration, pending the investigation of alternative vehicles of semantic content.

6. Concepts and inference

We cannot reach our conclusion directly from the claims that thoughts have semantic content, and that those contents are subject to the closure condition. But, fortunately, those two claims do not exhaust the significance of the neo-Fregean idea of conceptualised content.

It is a feature of the thought that a is F that entertaining that thought involves mastery of the constituent concept of being F; a piece of concept mastery that can be employed in further thoughts about other objects. So, it is not merely the case that if a thinker can think that a is F and think that b is G, then he can also think that b is F. It is not merely that there is one state of the thinker with the content that a is F and another state with the content that b is F. Rather, entertaining the thought that a is F and entertaining the thought that b is F involve the deployment of a common piece of concept mastery - mastery of the concept of being F - and a common piece of knowledge - knowledge what it is for something to be F.

This is part of what is involved in the idea of conceptualised content; but it is not captured by the closure condition, since that condition could be met by the occurrence of states that are quite autonomous. The closure condition would be satisfied provided that, whenever there are states with the contents that a is F and that b is G, there are also states - even states that are intrinsically quite unrelated - with the contents that a is G and that b is F.

If we take the claim about common pieces of concept mastery, and combine it with the familiar picture of thoughts related in an inferential web, then we can derive a consequence that is highly promising given the purposes of our argument. Indeed, this consequence is explicit in Evans (1981, p. 132):

> To have a belief requires one to appreciate its location in a network of beliefs. . . . To think of beliefs in this way forces us to think of them as structured states; the subject's appreciation of the inferential potential of one belief (e.g. the belief that a is F) at least partly depending upon the

> same general capacity as his appreciation of the inferential potential of others (e.g. the belief that b is F). . . . Possession of this general capacity is often spoken of as mastery of a concept.

A thinker who has the thought that a is F appreciates that from this thought it follows that a is H, say; and he also appreciates that from the thought that b is F it follows that b is H. But that is not all. It is not just that there is an input-output pattern in the inferences that the thinker is disposed to make. The two inferences are manifestations of a common underlying capacity; namely, mastery of the concept of being F.

As Evans himself makes clear, the notion of a capacity or disposition is not to be understood in terms of the bare truth of conditional statements, but rather in a 'full-blooded' way (1981, p. 329). The idea of a common capacity being manifested in the two inferences should be unpacked in terms of a common explanation, adverting to a common state (1982, p. 102). In short, there is causal systematicity relative to the input-output pattern in a thinker's inferential practice.

Here is a simple example. A thinker who has the thought that Bruce is a bachelor appreciates that from this thought it follows that Bruce is unmarried; he also appreciates that from the thought that Nigel is a bachelor it follows that Nigel is unmarried. The thinker appreciates the inferential potential of the two thoughts; and this depends in each case on the same general capacity, namely, mastery of the concept of being a bachelor.

In order to have either the thought that Bruce is a bachelor or the thought that Nigel is a bachelor, the thinker must grasp the concept of being a bachelor. This is a matter of knowing what it is for an object to be a bachelor; of knowing *inter alia* that to be a bachelor requires being unmarried. This single piece of knowledge - that for an arbitrary object to be a bachelor, that object must be unmarried - is implicated in both the inferential transitions that the thinker is disposed to make.

All this is just what is needed for the second stage of our argument. It is part of the neo-Fregean conception of a thinker that, in the arena of thought, there is a genuine causal systematicity of inferential transitions.

7. Two objections

Our argument for the LOT hypothesis is essentially complete. But we should pause to consider two objections that might be raised against the second stage of the argument. One concerns the appeal to Evans's work; the other begins from that very simple example of thoughts about being a bachelor.

Evans on the language of thought

It might be objected that there is something infelicitous about our reliance upon Evans. For Evans himself says (1982, pp. 100-101):

> It seems to me that there must be a sense in which thoughts are structured. . . . This might seem to lead immediately to the idea of a language of thought, . . . However, I certainly do not wish to be committed to the idea that having thoughts involves the subject's using, manipulating, or apprehending *symbols* - which would be entities with non-semantic as well as semantic properties, . . . I should prefer to explain the sense in which thoughts are structured, not in terms of their being composed of several distinct *elements*, but in terms of their being a complex of the exercise of several distinct conceptual *abilities*.

In this passage (just before the introduction of the Generality Constraint) Evans rejects outright a certain conception of the LOT hypothesis - according to which it involves *the*

subject's using symbols - and he denies that the idea of thoughts as structured leads *directly* to the LOT hypothesis.

But this presents no objection to our argument; on these two points, we can be in total agreement with Evans. First, it is certainly no part of the LOT hypothesis, as it is argued for here, that the conscious, thinking subject is presented with thoughts as entities with non-semantic properties. Indeed, that would arguably be regressive. The LOT hypothesis concerns the scientific psychological underpinnings of a subject's conscious mental life.

Second, our argument precisely does not move immediately to the LOT hypothesis from the idea of thoughts as structured. Rather, the argument follows Evans in moving first to the notion of the exercise of common capacities, and in construing capacities in a full-blooded way. The step from there to the LOT involves the conditional claim linking systematicity of process with syntactic structure in input states (established in the first stage). That step of the argument is not, apparently, anticipated by Evans; but nor is it considered and rejected by him.

Concept mastery and primitively compelling inferences

There are many things that a thinker might conclude about Bruce, given that Bruce is a bachelor, which he would not conclude about Nigel, given the thought that Nigel is a bachelor. A thinker might reasonably conclude that Bruce drinks a lot of Foster's Lager, while Nigel drinks a lot of Spanish champagne, for example. And where a thinker does draw similar conclusions about Bruce and Nigel, there is not, in general, any guarantee that the inferential transitions are products of a common capacity.

Consequently, the plausibility of the idea of causally systematic inferential transitions might seem to be an artifact of the definability of being a bachelor. So it might be objected that, for almost any concept other that the concept of being a bachelor, it would be implausible to insist that there is causal systematicity of inferential transitions.

This second objection might be coupled with the idea that, where mastery of a concept does not consist in knowing a definition, objects falling under the concept exhibit only a family resemblance. And that might be thought to undercut even further the idea of a common capacity being exercised in inferences concerning different objects.

However, we can respond to this objection by noting that mastery of a concept may be constituted by commitment to a set of inferential principles, without those principles amounting to a statement of necessary and sufficient conditions for application of the concept. An alternative development of the argument for causal systematicity of inferential transitions illustrates this point.

In recent work, Christopher Peacocke (1986, 1989a, to appear) has been articulating a theory of concept mastery. In *Thoughts* (1986), the idea is expressed in terms of the canonical grounds and the canonical commitments of certain classses of contents. In 'What are concepts?' (1989a), the idea is of a possession condition for a concept, where this is often a matter of a thinker finding certain patterns of inference primitively compelling. (For a brief account, see Peacocke 1989b.)

Here is an example that Peacocke (to appear) develops. Part of what is involved in mastery of the concept ***plus*** is finding this inferential transition (T) primitively compelling:

18 + 64 is a certain number n;

so

18 + (the successor of 64, *viz.* 65) is the successor of n.

Likewise, the master of ***plus*** finds this transition (T') primitively compelling:

11 + 23 is a certain number m;

so

11 + (the successor of 23, *viz.* 24) is the successor of m.

And, of course, there are indefinitely many other primitively compelling inferential transitions exhibiting the same pattern.

Now, on Peacocke's account, mastery of the concept ***plus*** involves more than that the thinker should find each of these transitions - or enough of these transitions - primitively compelling. What is required is that the thinker should find the transitions primitively compelling *in virtue of their form*.

This is not intended as the requirement that the thinker should be able to conceptualise or to formulate the form or pattern of inference (R):

Given: m + k = n;

Infer: m + S(k) = S(n).

Rather, the account is supposed to relate 'grasp of ***plus*** to the causal influence of a form of transition which is not necessarily conceptualized' (p. 000). The idea is that the form of transition is causally explanatory: it enters the causal explanations of particular transitions' being found primitively compelling. And this phenomenon - of causally explanatorily relevant forms or patterns - is one that can be found in humans and in machines.

In this alternative development of the second stage of our argument for the LOT, Peacocke's idea of inferences that are found primitively compelling in virtue of their form - of a causally influential form of transition - is glossed in terms of causal systematicity of process.

The proposal is that at least part of what is involved in particular inferences being found primitively compelling in virtue of their form is this. Mirroring the commonality in the inferences that are found primitively compelling - namely, their form - there should be a commonality in the causal processes that explain their being found so.

Given this gloss, and the close connection between causal systematicity and knowledge of rules, we might call the common state that figures in the causal explanations of the various particular inferences - such as (T) and (T') - a state of knowledge of rule (R). As Peacocke says, this knowledge of (R) does not require that (R) should be conceptualised or formulated by the thinker. Nor does it require that (R) should be explicitly represented in the thinker's cognitive machinery.

Provided that the state of knowledge of an inferential rule such as (R) functions to mediate actual transitions in thought - from the premise of (T) to the conclusion of (T), from the premise of (T') to the conclusion of (T'), and so on - we have here an alternative version of the second stage of our argument. It is a version that can subsume the simple cases such as that of the concept bachelor, without suggesting that its application is restricted to concepts that are definable.

Taken together with the conditional claim of the first stage, it requires that the input states of the transition mediator corresponding to an inferential rule should have a syntactic articulation. Thus it provides an alternative completion of our argument for the LOT hypothesis.

8. Connectionism, syntax, and systematicity

Given the argument that our commonsense conception of ourselves involves a commitment to the LOT hypothesis, we can now argue for a *prima facie* tension between that commonsense conception and the connectionist programme for modelling cognitive processes.

The argument turns upon the claim that typical connectionist networks do not exhibit causal systematicity of process, and syntactic structure in input states. Of course, connectionism comes in several varieties, and there are some networks that do have these features; examples can be provided by networks with local representation of all the primitive concepts of some classical task analysis. So let us be more specific. What is to be considered is connectionism with distributed representation. In particular, we focus on networks with microfeatural, dimension shifted, representation in the style of Smolensky (1988).

We can begin with the question whether connectionist networks have syntactically structured input states. Having in mind that syntax is relative to semantics, we should be explicit. The question is whether the input states of a network have syntactic structure relative to the standard or classical semantic description of what the network is doing. Are there properties of connectionist input states, as such, which line up with the primitive concepts used in a classical analysis of the task that the network is performing?

Syntax

If the representation in a connectionist network is distributed rather than local, then activation at an individual input unit cannot be regarded as the tokening of a syntactic element. The reason is simple; for to say that representation is distributed is just to say that individual units are not the vehicles of representation.

Thus, the input states of a network may be representing facts about coffee in various contexts: in cups and jugs, with or without sugar. But there is not an individual unit that represents the occurrence of coffee. The representation of coffee in a cup is not a matter of activation at a coffee unit and a cup unit. Rather, what represents coffee in a cup is a pattern of activation over many units.

However, this simple fact about distributed representation does not yet show that there is no syntactic description of connectionist input states. Activation at a single unit is just a limiting case of a subpattern of activation; and which units are included in a total pattern of input activation is certainly a determinant of the causal consequences of that state. So - given our minimal notion of syntactic property - a subpattern of activation over several units is the right sort of thing to count as the tokening of a primitive symbol, provided that the subpattern corresponds to a semantic property of the input states in which it occurs.

So, although there is no proprietary coffee unit, might there not be a specific distributed pattern of input activation that means coffee?

There certainly are networks that show subpatterns of input activation of the envisaged kind. In the networks examined by Ramsey, Stich and Garon (this volume), the input states that represent various propositions about dogs all share a common subpattern of activation over eight input units, given by the vector <11000011>. Similarly, the activation vector <11001100>, over those same eight units, occurs

whenever the proposition concerns cats, while whenever the proposition concerns having fur there is a common subpattern of activation over the remaining eight input units, given by the vector <00001111>. Consequently, the pattern of activation for the proposition that dogs have fur, for example, can be regarded as the tokening of two primitive symbols, with co-occurrence of subpatterns being the network's way of combining subject and predicate to make a sentence.

Nevertheless, what Smolensky says indicates that he does not see this as the typical case. Concerning the constituent subpatterns of activation that represent coffee in various contexts - coffee with sugar, coffee in a cup, coffee in a jug - Smolensky says (1988, p. 17):

> These constituent subpatterns representing coffee in varying contexts are activity vectors that are not identical, but possess a rich structure of commonalities and differences (a family resemblance, one might say).

(If we focus upon representations of coffee in a cup, coffee in a jar, and coffee on a tree, then we invite the response that, even by classical lights, different concepts are involved: coffee drink, coffee granules, and coffee beans. But the point about contextual variation of microfeatural representation carries over to the cases where that response is no longer plausible.) So, there simply is no strictly common subpattern of activation that can be identified as a syntactic element meaning coffee.

If Smolensky is right then, relative to a semantic description in terms of coffee, cups, jugs, and the like, the input states of a typical connectionist network with distributed representation will not have a syntactic description.

Essentially the same point could be made in terms of a network performing the sentence interpretation task sketched in Section 3. We can imagine that the input states that register which sentence has been uttered (or presented visually) make use of distributed microfeatural representation. And we can suppose further that the way in which the predicates, for example, are pronounced (or written) varies somewhat, depending upon the name with which they are combined. Consequently, it may be that the input representation of the occurrence of the predicate 'is drunk' varies between its occurrences in the contexts 'Martin is drunk', 'Andy is drunk' and 'Frank is drunk'. The constituent subpatterns may exhibit family resemblance, rather than identity. In that case, although the objects in the task domain have syntactic structure, the input states of the network will not (relative to their semantic description as representing the names, predicates, and sentences of the task domain).

Systematicity

This fact, that networks of a certain type do not have syntactically structured input states, need not threaten the conditional claim in the first stage of our argument for the LOT hypothesis. That conditional claim remains intact, so long as the networks under discussion do not exhibit causal systematicity of process (relative to input-output generalisations pitched at the level of semantic description in the terms of a classical task analysis). And, indeed, they do not.

Let us return to our hypothetical example of a network whose input states represent facts about coffee in various contexts. We can imagine that some of the input states mean, variously, that there is coffee in a cup, coffee in a jug, coffee in a glass, coffee with sugar, and some that there is wine in a cup, wine in a jug, wine in a glass, wine with sugar, and so

on. Similarly, we can imagine that some of the output states mean that there is warm drink in a cup, warm drink in a jug, warm drink in a glass, warm drink with sugar, and the like.

Now suppose that the network is performing some rudimentary inferential transitions. The input state with semantic content *coffee in a cup* produces the output state with semantic content *warm drink in a cup*; the input state with content *coffee with sugar* produces the output state with content *warm drink with sugar*; and so on. Viewing this from the outside, we can see a pattern in the input-output relation for this network: Whenever the input state means *coffee . . . Y*, the output state means *warm drink . . . Y*.

To ask whether the process that is going on in the network is causally systematic relative to that pattern is to ask whether the *coffee* to *warm drink* transitions all have a common explanation; whether there is, as a component of the network, a mechanism that is reponsible for all and only those transitions.

In general, the answer to this question is negative. It will not be strictly true that there is a common set of weights on connections that is implicated in all and only the *coffee* to *warm drink* transitions. In terms of knowledge of rules, we can say that it will not be correct to describe the network as having knowledge of the rule:

Given: there is coffee; Infer: there is warm drink

although, *ex hypothesi*, the network's behaviour conforms to that rule.

As in the discussion of syntax, we can make essentially the same point in terms of the sentence interpretation example. A network with distributed, microfeatural, input and output encoding may achieve conformity to the rule

Given: the sentence presented contains the predicate 'is drunk';

Infer: the proposition conveyed concerns the property of being drunk.

But, if the input representation of the occurrence of the predicate 'is drunk' varies from case to case, then the explanation of the network's conformity to this rule in one case will not be just the same as the explanation in another case.

None of this is to say, of course, that in connectionist networks completely distinct and autonomous processes are involved in the various transitions that accord with a pattern. Connectionist networks offer an option in between strict commonality and strict autonomy or modularity. They fall between systems with knowledge of rules, on the one hand, and mere look-up tables, on the other.

Connectionism presents no problem for the conditional claim that if there is causal systematicity of process then there is syntactic structure in the input states to that process. For, with distributed representation, there is typically neither syntax nor systematicity. What is more, there may not be systematicity of input-output process, even where there is syntactic structure in the input states.

This typical departure from causal systematicity does not, by itself, constitute an objection to the connectionist programme. If a cognitive process is causally systematic, then distributed connectionism is unlikely to provide a good model for that process. But it is, in general, an empirical question whether any given cognitive process is systematic in the sense that concerns us.

Consequently it is, in general, a matter for detailed empirical investigation whether modelling actual cognitive processes presents a problem for connectionism. However, the second stage of our argument reveals that there is a tension between the connectionist

programme for modelling cognition and our commonsense conception of ourselves as thinkers. On the face of it, the connectionist paradigm does not provide a good scientific psychological model for the domain of conceptualised thought and inference.

9. An invitation to eliminativism

If all this is right, then what seems to be in prospect is an argument from connectionism to eliminativism; not, to be sure, to the elimination of all semantic content, but to the elimination of the bearers of semantic content that belong in the commonsense scheme: beliefs, and thoughts in general.

The present argument for a tension between the commonsense scheme and the connectionist programme finds a parallel in the paper by Ramsey, Stich and Garon (this volume). They defend a conditional claim (p. 000):

> *If* connectionist hypotheses . . . turn out to be right, so too will eliminativism about propositional attitudes.

Their argument comes in two main stages. First, they claim that the commonsense scheme is committed to propositional modularity. This is the idea that (p. 000):

> propositional attitudes are *functionally discrete*, *semantically interpretable*, states that play a *causal role* in the production of other attitudes, and ultimately in the production of behaviour.

Then, second, they claim that distributed connectionist networks do not exhibit propositional modularity.

The argument of the present paper is likewise an argument for an incompatibility between a feature of the commonsense scheme and connectionist hypotheses.

Ramsey, Stich and Garon argue: Networks do not exhibit propositional modularity; the commonsense scheme is committed to propositional modularity; therefore connectionism is opposed to the commonsense scheme. Similarly, the argument of this paper runs: Networks do not exhibit syntax and causal systematicity of process; the commonsense scheme is committed to syntax and causal systematicity of process; therefore connectionism is opposed to the commonsense scheme.

The parallel extends to some points of detail. Ramsey, Stich and Garon argue that in a connectionist network there are no functionally autonomous vehicles of proposition-sized semantic contents. In the case where the putative vehicles under consideration are patterns of weights, their point is essentially similar to the claim that processing in networks is not causally systematic. This is hardly surprising. For suppose we focus upon the role of beliefs in mediating between desires and action, or in mediating inferentially between other beliefs. Then what propositional modularity requires is that there should be functionally autonomous transition mediators. And that is also what is required if the several transitions - from desire to action, or from belief to belief - are to be causally systematic.

Each argument purports to establish a necessary condition for a being to be a thinker (a believer, a deployer of concepts). In each case, this necessary condition concerns internal cognitive architecture; it is a condition that is far from being guaranteed by facts about behaviour. For any given being whose behaviour *prima facie* warrants the attribution to it of beliefs and other attitudes, in accordance with the intentional stance, it is a genuine epistemic possibility that the being does not meet the condition on internal architecture.

In each argument, connectionism serves to provide a vivid example upon which to focus what is a quite general issue. For, in each case, it is claimed that a being whose internal cognitive architecture is correctly described as a connectionist network will not meet the necessary condition for being a thinker which the argument purports to establish.

The general issue that connectionism brings so sharply into focus is this. Is it philosophically acceptable that an *a priori* argument should render it epistemically possible that *we* should turn out not to be believers or thinkers? One powerful source of resistance to our argument for the LOT hypothesis is precisely the thought that this is not acceptable; that it is built into our very conception of a believer or thinker that we are the paradigm exemplars. According to that view, the proposition that we are believers is philosophically non-negotiable.

In fact, it is more or less inevitable that philosophers who have any Wittgensteinian sympathies at all will feel some unease about our argument. Order might proceed, so to speak, out of chaos; and it might proceed out of order. It is an *a posteriori* matter which is the case. Part of the message of *Zettel* 608-9 is, perhaps, that philosophers have no business insisting that the system *must* 'continue further in the direction of the centre'. And the invitation to eliminativism then presents itself as the penalty for failing to heed that message.

However, despite the virulence of these doubts, we can fortify ourselves with two thoughts. First, it is possible to mount a defence against eliminativism without rejecting our argument. Second, blanket immunity against eliminativism is only purchased at an exorbitant price. These two claims will be defended briefly in the next (and final) section.

10. Defending belief

There are at least two broad ways of mounting a defence against eliminativism, while accepting our argument for the LOT hypothesis; but one can be dismissed quite rapidly.

The first way is to adopt an *a prioristic* stance towards the future of science. According to this first defensive strategy, we should allow that evidence might build up in favour of the hypothesis that our internal cognitive architecture does not meet the conditions which, according to the our argument, are necessary for being a believer. This is to say that it is conceivable that we should amass evidence such that, all else equal, the best explanation of that evidence would be that the LOT hypothesis is false. But in that situation we should then say that all else is not equal, and that in this case we have reason to maintain that what would otherwise be the best explanation of the evidence is not, in fact, the correct explanation.

If our argument for the LOT hypothesis had been an *a posteriori* argument, then this would be a viable strategy. Indeed, in the face of a strong *a posteriori* argument, the claim that evidence might pile up against the LOT hypothesis would appear question-begging. But given that the original argument is *a priori*, this first strategy is surely just an unjustifiable refusal to accept an inference to the best explanation.

So, in the context of an invitation to eliminativism issuing from an *a priori* argument, this first defensive strategy is not to be recommended.

The second defensive strategy against eliminativism involves a pincer movement.

For one component of the movement, we can return to *a posteriori* considerations in favour of the LOT hypothesis. These can be used to support the view that it is

empirically unlikely that the behaviour that we find could be reliably forthcoming without an internal architecture measuring up to the requirements of the LOT. Thus, *a posteriori* arguments for the LOT are not rendered dialectically redundant by our proposal for an *a priori* argument.

In fact, *a posteriori* arguments for the LOT can be divided into two types. There are some arguments which take the form of a 'How else?' challenge. As we have seen, in the context of a developing alternative paradigm such as connectionism, this type of argument is apt to seem question-begging.

But there are other arguments involving detailed evaluation of the performance of connectionist models that depart from the paradigm of rules and representations, systematicity and syntax. Suppose that analysis of the performance of networks were to uncover aspects which are attributable to the departure from systematicity and syntax, and which conspicuously differ from human performance. Then that would count against connectionism ever becoming the dominant paradigm for modelling human cognitive processes.

This idea of an aspect of a network's performance that is attributable to the departure from systematicity and syntax can be illustrated as follows.

The distinction between causally systematic processes and others is drawn in such a way that it can, in principle, be used to distinguish betwen two systems with the same input-output relation. However, in real cases, it is overwhelmingly likely that a departure from causal systematicity will show up somewhere in a system's input-output relation; particularly if we probe the system's operation by presenting novel inputs.

Thus - to use the familiar example of the past tense (Rumelhart and McClelland 1986) - suppose that the transitions from regular verbs to their past tenses have a common causal explanation: there is a common mechanism that mediates those several transitions. It follows that the input states for such verbs must have some common property (a symbol meaning that the verb is regular) to engage that common mechanism. If a new verb is presented then, provided that the input state has that same property (tokens the symbol meaning that the verb is regular), the verb will be awarded a past tense in just the same way as all other regular verbs.

The situation will be very different if there is only a family resemblance among the transitions for various regular verbs. In such a case, the family resemblance is dictated by similarities amongst input states, where those states are patterns of activation over units that individually respond to microfeatures of some kind. If a new verb is presented, then the transition to a past tense is conditioned by the microfeatural similarity of the new verb to others. If the new verb is microfeaturally very different from other regular verbs, then it is likely to be awarded a past tense in a very different way.

Thus, the highly deviant treatment of novel verbs that are microfeaturally remote from familiar examples - which is an aspect of the performance of the Rumelhart and McClelland network - is attributable to the departure from the rules and representations paradigm. If it turns out that human performance conspicuously differs from that of the network in this respect - as Pinker and Prince (1988) argue that it does - then that lends some non-question-begging support to the 'How else?' challenge.

So much for the first component of the pincer movement.

For the other component of the movement, we can point out that connectionist networks that do not *as such* employ syntactically structured vehicles of semantic

content are susceptible to analyses of their internal operation, such as cluster analysis or receptive field analysis. We can argue that, in some cases, these analyses vindicate higher levels of description at which we find a system that does meet the requirements of the LOT, even though the system is realised in a connectionist substructure.

In some other cases, the analyses reveal that the network can be regarded as composed of two devices. One is a front end recognition device that is connectionist through and through. The second is a device, which does - at some level of description - meet the requirements of the LOT, and which takes as inputs the outputs of the recognition network.

In short, the requirements of syntax and systematicity are typically not met at the level of description of networks in terms of units and connections, activation and weights. But that does not rule out the possibility that some analysis of the operation of a network may vindicate a higher level of description from whose point of view the approximate and blurred commonalities are just variable realisations of real commonalities. (See Clark 1989, 1990; Davies 1990b, 1990c.)

So much, very briefly, for the second component of the pincer movement.

If successful, this pincer movement renders it highly probable that we are actually believers; or, more accurately, renders it highly probable that we meet the particular necessary condition uncovered by our *a priori* argument.

Thus, the tension between the connectionist programme and the commonsense scheme can be reduced. But it is not altogether removed. For there is no absolute guarantee that, if we turn out to have connectionist networks inside our heads, then they will be networks that meet the requirements of syntax and systematicity (or of propositional modularity) at some vindicated level of description. We must live with the prospect that empirical discoveries about cognitive architecture may come into conflict with our commonsense conception of ourselves.

The dissatisfied critic

Suppose that someone insists that this second defensive strategy - the pincer movement - is insufficient to honour the intuition that our being exemplars of the property of being a believer is non-negotiable.

We could take a further step by acknowledging that what the critic regards as non-negotiable operates as a kind of presupposition of our practice in using the notions of a thinker, believer, or deployer of concepts. This would be to accept that these notions have no point for us unless they apply to us. But, if the critic is not satisfied with this presuppositional way of deferring to the intuition, then we have to argue that to go further in the direction that he wants would bring its own intolerable problems.

If it is to be non-negotiably true that we who produce interpretable behaviour are thinkers, then the concept of a thinker must impose no necessary conditions that go beyond behaviour. In particular, it must impose no necessary conditions at all upon internal cognitive architecture. But this means that what the critic wants is a form of behaviourism: not, to be sure, analytical behaviourism, but a doctrine that might be called *supervenient behaviourism*.

This form of behaviourism is itself arguably incompatible with the commonsense scheme. Imaginary examples of beings that produce the right behaviour by way of unusual internal architectures - the string-searching machine of Block (1981) or the Martian

marionette of Peacocke (1983) - reveal that supervenient behaviourism is out of line with our intuitions about thinkers. In any case, if the choice lies between behaviourism and facing up to eliminativism, then there are many of us who know which way we are voting.

The upshot is that the dissatisfied critic must remain dissatisfied. Blanket immunity against eliminativism is not to be purchased.

Conclusion

We began by setting aside some reservations about the very idea of a LOT. We then employed neo-Fregean resources to construct an *a priori* argument for the LOT hypothesis - an argument that proceeds in two main stages. This argument has the consequence that there is a *prima facie* tension between the connectionist programme and our commonsense conception of ourselves as thinkers.

The prospect then opens up of an argument from connectionism to eliminativism; and that prospect is a potential source of resistance to the argument for the LOT hypothesis. It is possible to defend the commonsense scheme, and to go some way towards honouring the intuition of its non-negotiability. But we should resist pressure to empty our conception of ourselves of all causal commitments. Rather, we have to face up to the possibility that developments in scientific psychology may oblige us to revise that conception more or less drastically.[1]

Note

1 Thanks to Katherine Morris for comments on an early version read to the Oxford Philosophical Society in November 1988; and to Ned Block, Andy Clark, and Christopher Peacocke for countless conversations on these topics. Talks based on this material were given at the Australian National University, the University of Sydney, the University of Queensland, and LaTrobe University, during August and September 1989, and at MIT and Rutgers University, during February and March 1990. I am grateful to ANU and the British Academy for financial support.

References

Barwise, J. 1987 Unburdening the Language of Thought, *Mind and Language* vol.2, pp. 82-96

Block, N. 1981 Psychologism and Behaviorism, *Philosophical Review* vol.90, pp. 5-43

Churchland, P.M. 1986 Reductive Strategies in Cognitive Neurobiology, *Mind* vol.95; reprinted in *A Neurocomputational Perspective*, Cambridge, MA: MIT Press, pp. 77-110

Clark, A. 1989 Beyond Eliminativism, *Mind and Language* vol.4, pp. 251-79

Clark, A. 1990 Connectionist Minds, *Proceedings of the Aristotelian Society* vol.90, pp. 83-102

Crane, T. 1990 The Language of Thought: No Syntax Without Semantics, *Mind and Language* vol.5, pp. 187-212

Davidson, D. 1973 Radical Interpretation. In *Inquiries into Truth and Interpretation*, Oxford: Oxford University Press, 1984, pp. 125-39

Davies, M. 1987 Tacit Knowledge and Semantic Theory: Can a Five Per Cent Difference Matter? *Mind* vol. 96, pp. 441-62

Davies, M. 1989 Tacit Knowledge and Subdoxastic States. In A. George (ed.), *Reflections on Chomsky*, Oxford: Blackwell, pp. 131-52

Davies, M. 1990a Thinking Persons and Cognitive Science, *AI and Society* vol.4, pp. 39-50

Davies, M. 1990b Rules and Competence in Connectionist Networks. In J. Tiles (ed.), *Evolving Knowledge in Natural Science and Artificial Intelligence*, London: Pitman, pp. 85-114

Davies, M. 1990c Knowledge of Rules in Connectionist Networks, *Intellectica* no.9: D. Memmi and Y.M. Visetti (eds.), ***Connectionist models***, pp. 81-126

Dennett, D. 1971 Intentional Systems. In *Brainstorms*, Montgomery VT: Bradford Books, 1978, pp. 3-22

Dennett, D. 1981 True Believers. In *The Intentional Stance*, Cambridge MA: MIT Press, 1987, pp. 13-35

Devitt, M. 1989 A Narrow Representational Theory of the Mind. In S. Silvers (ed.), *Rerepresentation: Readings in the Philosophy of Mental Representation*, Dordrecht: Kluwer Academic Publishers, pp. 369-402; reprinted in W.G. Lycan (ed.), *Mind and Cognition: A Reader*, Oxford: Blackwell, 1990, pp. 371-98

Evans, G. 1981 Semantic Theory and Tacit Knowledge. In S. Holtzman and C. Leich (eds.), *Wittgenstein: To Follow a Rule*, London: Routledge and Kegan Paul; reprinted in Evans, G. *Collected Papers*, Oxford: Oxford University Press, 1985, pp. 321-442

Evans, G. 1982 *The Varieties of Reference*, Oxford: Oxford University Press

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

Fodor, J. 1985 Fodor's Guide to Mental Representation, *Mind* vol. 94, pp. 77-100

Fodor, J. 1987a *Psychosemantics*, Cambridge, MA: MIT Press

Fodor, J. 1987b A Situated Grandmother? *Mind and Language* vol. 2, pp. 64-81

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

Hinton, G.E., McClelland, J.L. and Rumelhart, D.E. 1986 Distributed Representations. In D.E. Rumelhart, J.L. McClelland and the PDP Research Group, *Parallel*

Distributed Processing: Explorations in the Microstructure of Cognition Volume 1: Foundations, Cambridge, MA: MIT Press, pp. 77-109

Peacocke, C. 1983 *Sense and Content*, Oxford: Oxford University Press

Peacocke, C. 1986 *Thoughts: An Essay on Content*, Oxford: Blackwell

Peacocke, C. 1989a What Are Concepts? In P.A. French, T.E. Uehling and H.K. Wettstein (eds.), *Midwest Studies in Philosophy Volume 14: Contemporary Perspectives in the Philosophy of Language II*, Notre Dame, IN: University of Notre Dame Press, pp. 1-28

Peacocke, C. 1989b Possession Conditions: A Focal Point for Theories of Concepts, *Mind and Language* vol. 4, pp. 51-6

Peacocke, C. to appear Content and Norms in a Natural World. In E. Villaneuva and L. Valdivia (eds.), *Information-Theoretic Semantics and Epistemology*, Oxford: Blackwell

Perry, J. 1986 Thought Without Representation, *The Aristotelian Society Supplementary Volume 60*, pp. 137-51

Pinker, S. and Prince, A. 1988 On Language and Connectionism: Analysis of a Parallel Distributed Processing Model of Language Acquisition, *Cognition* vol. 28, pp.73-193

Rumelhart, D.E. and McClelland, J.L. 1986 On Learning the Past Tenses of English Verbs. In J.L. McClelland, D.E. Rumelhart and the PDP Research Group, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition Volume 2: Psychological and Biological Models*, Cambridge, MA: MIT Press, pp. 216-71

Smolensky, P. 1987 The Constituent Structure of Connectionist Mental States, *The Southern Journal of Philosophy* vol.26 Supplement, pp. 137-61

Smolensky, P. 1988 On the Proper Treatment of Connectionism, *Behavioral and Brain Sciences* vol.11, pp. 1-74

Stich, S. 1978 Beliefs and Subdoxastic States, *Philosophy of Science* vol.45, pp. 499-518

Stich, S. 1983 *From Folk Psychology to Cognitive Science*, Cambridge, MA: MIT Press

Wittgenstein, L. 1969 *The Blue and Brown Books*, Oxford: Blackwell

Wittgenstein, L. 1981 *Zettel*, Oxford: Blackwell