# CONSCIOUSNESS, CONTENT, AND COGNITIVE ARCHITECTURE

by

MICHAEL VERNE ANTONY

B.Sc., University of Toronto
(1985)

SUBMITTED TO THE DEPARTMENT OF
LINGUISTICS AND PHILOSOPHY
IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR
THE DEGREE OF

DOCTOR OF PHILOSOPHY

in

PHILOSOPHY

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

October, 1990



Signature of Author_____
Department of Linguistics and Philosophy
October, 1990

Certified by_____
Ned Block, Professor of Philosophy
Thesis Supervisor

Certified by_____
Robert Stalnaker, Professor of Philosophy
Thesis Supervisor

Accepted by_____
George Boolos, Professor of Philosophy
Chair, Committee on Graduate Studies

## CONTENTS

Essay 3

# CONSCIOUSNESS, CONTENT, AND COGNITIVE ARCHITECTURE

by

Michael Verne Antony

**Abstract**

This thesis consists of three essays in the philosophy of mind. Essay 1 contains an argument against functionalist theories of consciousness. The argument exploits an intuition to the effect that parts of an individual's brain (or of whatever else might realize the individual's mental states, processes, etc.) that are not *in use* at a time t, can have no bearing whatever on whether that individual is conscious at t. After presenting the argument, I defend it against two possible objections, and then distinguish it from two arguments which appear, on the surface, to be similar to the argument of this essay.

Essay 2 takes up Tyler Burge's thesis, based on his thought experiments in 'Individualism and the Mental', that propositional attitudes are properties individuals have in virtue of social relations they bear to other language users. An alternative interpretation of Burge's thought experiments is offered on which the intuitions Burge evokes can be accepted while his conclusions about the social character of thought are denied. The alternative interpretation given, I go on to argue that it is preferable to Burge's.

Essay 3 concerns Fodor and Pylyshyn's argument against connectionism as a theory of the cognitive architecture. That argument contains two premises: first, that the Language of Thought is true at the level of cognitive architecture, and second, that connectionism and the Language of Thought are incompatible. I argue that Fodor and Pylyshyn's defenses of both premises fail, and I provide positive reasons for supposing the second premise is false.

Thesis Supervisors: Ned Block, Robert Stalnaker,
Professors of Philosophy

## ACKNOWLEDGEMENTS

Heartfelt thanks to each of the following:

Ned Block and Bob Stalnaker, my thesis supervisors, for extensive comments on many drafts of each paper in the thesis, and for several other ways in which they have generously given of their time.

Rob Cummins, Martin Davies, Eric Lormand, Paul Pietroski, Georges Rey, Gabe Segal, and Stephen White, for written comments on various parts of the thesis, long discussions on such parts, or both.

Lynd Forguson, Bob Lockhart, and Doug Creelman, three of my teachers from the University of Toronto, for all their assistance and encouragement over the years.

The Social Sciences and Humanities Research Council of Canada for funding part of this work.

Ann Bumpus, Kate Kearns, Marga Reimer, Brian Ulicny, Catherine Womack, and others in the department at MIT, for friendship, support, dude talk, squash, etc.

John Minahan for stimulating talks on mountains, Pat Goebel for entertaining trans-continental e-communication, Lisa Ryan for her friendship and love, and my parents and the rest of my family for things too numerous to list.

## Essay 1

# AGAINST FUNCTIONALIST THEORIES OF CONSCIOUSNESS

In this essay I offer an argument against functionalist theories of consciousness. Like more familiar arguments on this topic[1], it depends crucially upon pretheoretical intuitions about consciousness. However the intuition driving the argument is novel: it concerns the idea that conscious experiences are intrinsic with respect to time - in particular, that they do not depend for their identity upon bearing causal relations to events that occur after them.

I proceed as follows. In the first section I state and explicate the argument. In section 2, I discuss the two main lines of response that are open to the functionalist. And in the final section I compare and contrast the argument from Section 1 with two related arguments in the literature.

## 1 The Argument

The argument of this essay is directed at the metaphysical thesis that conscious, experiential, or phenomenal kinds are functional kinds. And that I take as the simple and unadorned claim that what makes a conscious kind the kind it is

---

[1] Five of the best known are these: (1) The absent qualia argument (as in Block (1978)); (2) The inverted spectrum argument (as in Shoemaker (1981a)); (3) Jackson's (1982) "knowledge argument"; (4) Nagel's (1974) argument from "What Is It Like to Be a Bat?"; (5) Searle's (1980) "Chinese Room" argument. A sixth argument, similar to my own, has recently been advanced by Tim Maudlin (1989). I discuss Maudlin's argument in section 3.

its particular causal relations to inputs (e.g., environmental stimuli), outputs (e.g., bodily motions), and other mental kinds.[2] In general, according to functionalism, a conscious kind Q is what it is in virtue of being causally related to some set of mental or environmental kinds, $\mathbf{C} = \{C_1, C_2, ..., C_m\}$, which are direct or indirect causes of Q, and some set of such kinds, $\mathbf{E} = \{E_1, E_2, ..., E_n\}$, which are direct or indirect effects of Q.[3]

Let Q be any experiential, phenomenal, or conscious kind whatever. Q might be a pain event, an emotional state, a phenomenal tactile experience, a conscious visual process, and so on. If Q is an activated or occurrent kind, a certain internal structure, and perhaps a minimal duration, may be essential to its identity. Suppose now that in the brain of some arbitrary subject, who I shall dub 'Sam', Q is currently instantiated. And imagine that under the current conditions, Q

---

[2] I employ 'mental kind' as a general term to subsume mental states, mental processes, mental events, and the like. Accordingly, I use 'mental kind' (and 'conscious kind', etc.) to refer to either tokens or types, depending on the situation (just as is common with the more specific terminology). Context determines the appropriate interpretation in each case.

[3] Two points: First, what a functional kind causes at a given time will in general depend upon what other conditions obtain at that time. What action a belief causes depends on one's desires; what the effects are of a Turing Machine in state S depends on what is on the tape, and so forth. Thus each element of **E** should be taken to be a "conditional effect" of Q, i.e., one produced only under certain conditions. And because combinations of environmental and mental kinds often will be necessary for causing Q, the elements of **C** should be assumed to include such combinations. (See Shoemaker (1981b) for a related discussion.)

Second, though I shall continue to speak of the elements of **C** and **E** as the causes and effects of Q, strictly speaking, the relations between Q and the elements of **C** and **E** need not be causal, just counterfactual-supporting. The conditional 'Q -> E', for instance, need only be supported by the counterfactual 'If Q had not been instantiated, neither would have E'. Such support can be had if there is a third factor causing both Q and E in such a way that E is caused only if Q already has been.

is disposed to cause some arbitrary element from the set, **E**, of Q's effects. Call that element 'E'. The argument, then, runs as follows:

(1) Q is a functional kind (assumption). Therefore the causal relatedness of Q to E is essential to the identity of Q.

(2) Q is instantiated in Sam's brain before E. That is, Q is instantiated at some time $t_1$, and then at $t_2$, where $t_2$ is later than $t_1$, E is instantiated.

(3) Between $t_1$ and $t_2$, while whatever realizes Q in Sam's brain remains instantiated, it is possible to make it such that E cannot be instantiated. This can be done, for example, by destroying the relevant region of Sam's brain.[4]

(4) Doing what is said to be possible in (3) does not stop Q from being instantiated in Sam, at least during the interval between $t_1$ and $t_2$. For how could destroying an *unused* brain part have any effect at all on Sam's phenomenal experience between $t_1$ and $t_2$?

---

(5) It follows that Q is not a functional kind. For by (4), being causally related to E is not essential to the identity of Q (at least between $t_1$ and $t_2$), contrary to the assumption in (1).

---

[4] I assume Q and E are realized in distinct regions of Sam's brain; and also that the region that realizes E serves no other function. I discuss these assumptions below.

The above argument appears to be a valid instance of *reductio ad absurdum*: on the assumption that Q is a functional kind, it follows both that Q must be causally related to E and that it need not be. Therefore Q is not a functional kind. What needs to be defended, then, are Premises (2), (3), and (4). I consider them in turn.

### *1.1 Premise (2)*

Premise (2) I take to be relatively uncontroversial. If Q *causes* E, then Q *precedes* E, unless effects can occur before or simultaneously with their causes. I assume effects cannot occur before their causes. And though one might argue that, in certain cases, causes occur simultaneously with their effects, clearly the bulk of causal relations specified by the functionalist are not of that sort. Let Q and E, and the relation between them, therefore, constitute a standard case.

### *1.2 Premise (3)*

> (3) Between $t_1$ and $t_2$, while whatever realizes Q in Sam's brain remains instantiated, it is possible to make it such that E cannot be instantiated.

Suppose Q and E are realized by neural activity in two distinct regions of Sam's brain, say $R_Q$ and $R_E$. This, of course, is not to beg any questions and say that activity in these regions is *sufficient* for the instantiation of Q or E; only that

activation of those regions realizes Q and E when all other necessary conditions are satisfied, whatever they may be. Assume that Q causes E by way of neural tissue that connects $R_Q$ to $R_E$.[5] Then if the tissue constituting $R_E$ is destroyed between $t_1$ and $t_2$ while $R_Q$ is still activated, then for some time after the destruction, what realizes Q in Sam will be instantiated while it is impossible for E to be instantiated. Thus, Premise (3).

In my defence of (3), I have assumed that Q and E are realized in distinct regions of Sam's brain, thus enabling one region to remain intact while the other is destroyed. I also have assumed that $R_E$'s only function is to realize E when activated.[6] One might object, however, that those assumptions cannot freely be made, since either Q and E, or E and some other kind, might be realized in the *same hardware*, e.g., in distinct activation patterns over the same set of neurons.

Well, of course, that is possible. But it is not necessary. Functionalism explicitly allows for alternative ways of realizing the same theory. Indeed, that is one of its chief advertised strengths. Thus we would be led to believe that any functionalist theory, all of whose kinds are realizable in the same hardware, is also a theory whose kinds are realizable in distinct bits of hardware. True enough, functionalists often point out that certain realizations of a theory, so far as we know, might be nomological impossibilities. But that surely can be of no help here: there is no reason to think the class of cases I have imagined are impossible. In

---

[5] If 'Q --> E' is merely a counterfactual-supporting conditional, assume some distinct causal factor causes activation first in $R_Q$, and then in $R_E$, in such a way that $R_E$ becomes activated only if $R_Q$ already has.

[6] See fn. 5.

my defence of Premise (3), therefore, I assume one possible realization of the theory that includes Q and E. If functionalism fails with respect to that realization of that theory, functionalism fails with respect to that theory.

### *1.3 Premise (4)*

Obviously (4) is the controversial premise. In essence, (4) states that even if $R_E$ is destroyed while $R_Q$ remains activated, so that it is impossible for Q to cause E, nevertheless Q is still instantiated in Sam - at least until $t_2$ when E would have been instantiated had no damage been done.

To make this vivid, imagine that Q is instantiated in Sam's brain at $t_1$. Starting at $t_1$, therefore, Sam begins to undergo a conscious experience, or enters some phenomenal state such as pain, say. Now the instantiation at $t_1$ of that pain (Q) coincides with the activation of the region of Sam's brain, $R_Q$. Suppose that it takes one second for a signal to travel from $R_Q$ to $R_E$ - the region of Sam's brain that, in these circumstances, realizes the effect (E) that the pain is disposed to cause. (E, we might imagine, is the construction of a plan to take an aspirin.) Let $t_2$ be the time $R_E$ first gets activated. Finally, imagine $R_Q$ remains activated throughout the interval from $t_1$ to $t_2$.

Now the claim of Premise (4) is this: If Sam is in pain between $t_1$ and $t_2$ when $R_Q$ is activated and $R_E$ is *sitting idle* (hence playing no role in instantiating any mental kind whatever), then Sam is in pain between $t_1$ and $t_2$ when $R_Q$ is activated and $R_E$ is *destroyed.* To express the intuition bluntly, merely tampering

with some distant, unused part of Sam's brain between $t_1$ and $t_2$ can have no bearing at all on whether Sam is experiencing pain, or any other conscious kind during that interval.[7]

I think the point is brought out more strikingly in the following way (though doing so is inessential to the argument). Imagine there exists some impressive technology that enables one to incapacitate instantly any region of Sam's brain one chooses, and then just as instantly restore that region to its normal state. Then at some time between $t_1$ and $t_2$, when the neural activity is propagating from $R_Q$ to $R_E$, one could incapacitate the inactive region, $R_E$, for 250 msec., say, and then restore it to normalcy - all prior to $t_2$. At $t_2$, when the signal reaches $R_E$, $R_E$ would become activated, and E would be realized as though nothing unusual had happened. Now Premise (4) entails that the quarter-second incapacitation of the inactive region $R_E$ makes no difference to whether Sam is in pain (realizes Q) at any time between $t_1$ and $t_2$. For it just seems intuitively false that Sam's pain would cease during that interval, and then resume at its end. Fiddling with idle brain-parts, one would have thought, cannot be relevant to one's phenomenology.[8]

---

[7] In speaking of a brain region as being "unused" at t, I mean only that it is not serving to realize any mental kinds at t in the following narrow sense: at t, it is not realizing or instantiating *the occupants of any functional roles* specified by the psychological theory.

[8] Ned Block has suggested the following worry about this second way of defending Premise (4). It might be thought that there is some sleight of hand involved. Specifically, one might suppose that because $R_E$ will be restored to normalcy before $t_2$ by the operator of the "brain-machine," Q *is* disposed to cause E during the 250 msec. when $R_E$ is incapacitated. And that is because the operator has the intention to restore $R_E$ before $t_2$. No wonder, then, it seems Sam's pain would not be affected by the switching off and on of the not-yet-activated $R_E$. Consequently, to maintain that the causal relation between Q and E is broken during the incapacitation of $R_E$ is just a bit of trickery.

*1.4 Conclusion (5)*

Given Premise (4), and assumption (1), it follows that Q is not a functional kind, that it has no correct functional account, no true functional definition. And it should be clear that this conclusion holds regardless of whether Q is a state, event, or process - so long as Q is a conscious or experiential kind. For no part of the argument involved any important distinctions along these lines. Now since 'Q' is just an arbitrary place-holder for the name of any experiential or conscious kind, the above argument holds of all such kinds, and thus amounts to a general argument against functionalism with respect to consciousness.

## 2 Two Rejoinders

It seems to me there are two main lines of response open to the functionalist. The first is based on claims such as the following: functional definitions of mental kinds are *idealizations*; they merely point to *tendencies* to cause and be caused; the conditions they specify are meant to apply only under

---

But I think this is a dangerous route to take. In general, the functionalist should not want events outside an individual's body to determine whether mental kinds within that individual's brain are causally related. For example, the possibility that Sam's pain might start and stop a number of times solely because the brain-machine operator continually changes his mind about whether to restore $R_B$, surely should be excluded. Such a possibility, however, is precisely what taking the above line allows. Accordingly, it must be held that the causal connection between Q and E *is* severed between $t_1$ and $t_2$, even though the operator has every intention to restore $R_B$ at $t_2$.

*normal* circumstances, or where "other things are equal;" and so on. The second response is more direct. It involves trying to argue that the intuition expressed in Premise (4) is simply an illusion, on par with the intuition that speeds faster than light must be attainable.

### *2.1 Idealizations*

In the first section, I omitted from my characterization of functionalism the idea that the causal relations specified in the definition of a kind are those the kind exhibits under *ideal conditions*. However many functionalist theories appeal to this condition in one form or other. For example, in accounts like Lycan's (1987), and the more recent views of Dennett (1987), a mental kind's ideal causal relations are the ones it evolved to enter into. In Lewis' (1980) version of functionalism, the ideal relations a mental kind has with respect to an entity, S, are cashed out in terms of what is normal or typical for the population S is a member of. Other accounts, like Dennett's (1978) for instance, simply appeal to normal conditions without saying what that comes to. And the notion of a *tendency to causally interact* is also often invoked in the absence of any discussion of what it is to have a tendency.

Now it might seem inappropriate that reference to ideal conditions was excluded from my portrayal of functionalism. For many are likely to hold that it is that very feature of functionalism that renders my argument unsound. They might respond,

> Of course Q is instantiated when E is disabled, when Q no longer in fact bears the causal relations mentioned in Q's definition. That is because Q *normally* bears those relations, because Q bears those relations *ceteris paribus*, because Q *tends* to bear them, etc. That is enough for the activation of $R_Q$ to instantiate Q, even when some of Q's essential causal connections no longer obtain. Consequently, the conclusion and intuition of the argument can be accepted without harm to functionalism. All that must be given up is the version of functionalism described in section 1. But no one believed that anyway. So it would appear the argument does not work after all.

I left out reference to ideal conditions mainly so as not to complicate the presentation of the argument by dealing with that issue concurrently. As I shall now argue, however, it is difficult to see how idealizations can save the functionalist. For in all cases in which an account of ideal conditions has been provided, the move has the following unsavory consequence: it requires the functionalist to deny that consciousness supervenes on the physical make-ups of individuals.

Here is why. The functionalist who refers to ideal conditions claims that when $R_Q$ is activated and $R_E$ is destroyed, Q continues to be instantiated in Sam. And that is because, under the ideal conditions, or according to the tendency specified by the theorist, Q does cause E. Now to show that taking this line requires one to give up supervenience of the experiential on the physical, I help myself to a little science fiction (that is nevertheless consistent with quantum physics). I imagine an entity that instantly materializes from the surrounding matter

into a state physically identical to the state Sam is in when $R_B$ is first destroyed. Call this fellow 'Sam*'.

The reason supervenience must be given up is plain. On all existing accounts of idealization, there is no basis for saying that in Sam*, activation of $R_Q$ causes activation of $R_B$ under ideal conditions, that it tends to cause it, etc. I discuss these accounts presently. For now, however, notice that it must be specified what "normal" or "ideal" comes to, or what is meant by "other things being equal," or by a "tendency." For in the absence of such details, the functionalist can be charged with having begged the question. The trouble is that it is all too easy for ideal or normal conditions to amount simply to *those conditions under which the theory gets things right.* Similarly with the notion of a tendency: the nonoccurrence of any causal interaction specified in a functional definition can be interpreted as confirming either the presence or absence of the corresponding tendency. But, of course, *any* functional definition can be saved in that way. The functionalist, therefore, must demonstrate that the notion of idealization appealed to is a more substantive one.

Consider now those accounts of ideal conditions that have been fleshed out to some degree. Take evolutionary accounts. These fail because, in Sam*, activity in $R_Q$ did not evolve for the purpose of causing activation of $R_B$. Sam* did not evolve at all. And consider Lewis' (1980) story, according to which ideal conditions for Sam* are determined by what is normal in Sam*'s population. That is also inadequate since it is false that activity in $R_Q$ typically causes activation of $R_B$ in Sam*'s population. There is no such population. (Or at least Sam* belongs

to indefinitely many populations, if any.) And, of course, any account that cashes out the notions of tendency or normalcy in terms of frequency must also fail, since activation of $R_Q$ in Sam* has never caused $R_E$ to become activated. In all of these cases therefore - as far as I know they exhaust what is in the literature - the functionalist has no grounds for claiming that Sam* is undergoing a conscious experience. But if Sam* is not, and Sam is, supervenience of the conscious on the physical constitution of the individual must be rejected.[9]

This result should be difficult for any materialist to accept. Phenomenal experience does not seem to be the sort of thing that could be relational in that way. Though similar arguments have been marshalled against evolutionary theories of the *semantic content* of mental kinds, and though theorists[10] have been willing to accept that the analog of Sam* lacks thoughts with semantic content, that move appears not to be available in the case of consciousness. That having been said, I confess I have no argument that supervenience must be maintained.

There is a reason, however, why it would be extremely odd in this context for the functionalist to take that route.[11] The intuition expressed in Premise (4) can be put roughly as follows: Consciousness supervenes "locally" on its physical realization. By appealing to idealizations, one in effect accepts this claim of local supervenience. But why accept local supervenience only then to give up

---

[9] In general, Sam* shows that supervenience must be given up whenever ideal conditions for an individual are determined by viewing that individual in relation to a particular environment or history. For Sam* can always lack those relations.

[10] See Millikan (1984), for example.

[11] I owe the following observation to Bob Stalnaker.

supervenience on the individual? Why suppose unused parts of the brain are irrelevant, and maintain, e.g., that which population one is a member of is relevant? One gets more - not less - relationality in that way. One would do better, it would seem, to reject the intuition of local supervenience as illusory (a response I consider presently), and preserve supervenience upon the individual's physical make-up.

In any event, regardless of how one feels about holding on to supervenience on the individual's physical constitution, the following point holds: *If* one wants to appeal to idealizations, *and* one wants to hold on to supervenience on the individual's physical make-up, then a new substantive account of idealization is needed - one that gives the same results for Sam and Sam* - in order to avoid the charge of having begged the question. Without such an account, simple appeals to ideal conditions are subject to that charge. Certainly it is possible that such an account might be developed, in which case it can be employed to respond to the argument. In the mean time, however, the functionalist must look elsewhere.

### *2.2 The Intuition Is an Illusion*

The second way of responding to the argument is to claim that the intuition expressed in Premise (4), to the extent that one has it at all, is an illusion. True enough, it seems as though a conscious experience occurring at t could in no way be affected by tampering with brain regions not in use at t. But things are not always as they seem. I can think of two ways one might try to motivate this claim.

First, one might appeal directly to the standard sorts of functionalist intuitions, and claim that they provide reason for doubting the reliability of the intuition expressed in (4). A functionalist could be imagined to argue as follows:

> It might *seem* that Sam's pain (Q) would still be realized between $t_1$ and $t_2$ when $R_E$ is incapacitated, but consider what that implies. Depending upon what E is, it could mean that while $R_E$ is inoperative, there is no disposition in Sam to express any discomfort whatever; or to say 'yes' when asked 'Are you feeling any pain?'; or to plan actions with the aim of relieving pain; and so forth. But anyone who is not disposed to express pain, think about relieving oneself of pain, acknowledge being in pain when asked, etc., *is not in pain.* Appearances to the contrary, therefore, during the time $R_E$ is incapacitated, and the disposition essential to the realization of pain is no longer in place, Sam is not in pain.

How do these functionalist reflections acquire their force? Well, the functionalist asserts that particular dispositions are constitutive of pain. In evaluating that, we imagine a person - Sam, let us say - who lacks those dispositions. And that we do, for example, by picturing Sam, just after he has been asked if he is in pain, sitting comfortably in his chair with a contented look on his face, saying, 'No, not at all. Why do you ask?' Having conceived Sam in that way, we then conclude that of course he is not in pain.

But let us examine the matter more closely. Notice that engaging in this exercise of imagination and judgment does not, in itself, tell us whether pain is a dispositional (functional) property, instead of a more intrinsic one. For there are at

least two reasons one might withhold attributing pain to Sam in the above story. First, given Sam's appearance, given what he says, and given we are justified in believing he is not displaying a brilliant piece of acting, it would be *absurd* to say he is in pain. Moreover, it is hard to imagine what reasons we could have for attributing pain to Sam under those circumstances. The second reason for not ascribing pain to Sam, however, is rather different. It might be thought that to be disposed to speak, think, and act as Sam does *just is* not to be in pain. Here the question how it could be possible for Sam to be in pain, and yet behave as he does, would not even arise; just as the question would never arise whether a substance, known to be harmless to humans, might nevertheless be poisonous to humans. The first reason, therefore, is more of an epistemic one; the second, more conceptual or semantic.

If what I have said is correct, it follows that one will be convinced by the functionalist's reflections about Sam only if one *already* thinks pain is a dispositional property. If one does not, one might be baffled, and at a loss for what to say about how Sam could be in pain. But after all, one might reflect, the neuropsychological literature is chock-full of puzzling psychological phenomena that would have been thought impossible before their discovery. The functionalist's considerations, in themselves, therefore, are insufficient to force the "anti-functionalist" to abandon his intuitions about consciousness.

However consider the intuition underlying (4). If the functionalist agrees that it *seems* that tampering with idle brain parts is irrelevant to one's phenomenology, the functionalist expresses an anti-functionalist intuition, plain and simple. The

functionalist cannot say, 'I agree that incapacitating unused brain regions is ineffectual, but that is consistent with functionalist intuitions about consciousness for such and such reasons.'[12] If the functionalist could say that, there would be a stand-off between conflicting intuitions, and there would be reason to be wary of both. But the functionalist cannot. The anti-functionalist, on the other hand, can accommodate the functionalist's reflections about Sam, as we have seen. But the functionalist is not in a comparable position. Accordingly, I think it must be concluded that the standard functionalist considerations provide no reason for thinking the intuition expressed in (4) is illusory.

The second way of attempting to motivate the claim that the intuition underlying (4) is an illusion is as follows. One might argue that the intuition is not simply an intuition about the nature of conscious kinds - about pains, itches, and the like - but, in addition, it depends implicitly upon accepting some such notion as *qualia* as a coherent one. But that notion is fraught with confusion and inconsistency. No wonder, then, it leads one falsely to believe (4).

To support the claim that the notion of qualia is confused, one might look to a discussion of Dennett's (1988) for help. Dennett describes two coffee tasters, Chase and Sanborn, both of whom used to like the taste of Maxwell House coffee, but no longer do. Chase's explanation for his changed attitude is that he has become a more sophisticated coffee drinker, and no longer enjoys Maxwell House. The idea behind Chase's thought is that he has adopted a new attitude toward a

---

[12] Of course, the functionalist can say something like that by appealing to ideal conditions. But that is a different response, and the functionalist cannot have it both ways.

certain quale, in particular, toward the taste of Maxwell House. Sanborn's explanation for his own changed behavior is that something has changed in his "taste-analyzing perceptual machinery," and consequently Maxwell House no longer tastes the way it used to. Here, the quale *itself* is supposed to have changed, rather than merely Sanborn's qualia-directed attitudes. Dennett argues thus: There is no coherent notion of qualia that can support the distinction between Chase's and Sanborn's explanations of their own changed attitudes toward Maxwell House. But supporting such distinctions, supposedly, is precisely what a notion of qualia is needed for. Therefore there is no reason to think qualia exist.

I shall not delve into Dennett's argument any further here. Though I think it has difficulties, and certainly does not support the nonexistence of qualia, nevertheless it does seem to me to suggest, at least, that intuitions about qualia are not as robust as they might appear. And that, in itself, might be taken as reason enough to adopt a skeptical stance toward the intuition expressed in (4) - to suspect it might be an illusion.

But I think that would be a mistake, and for two reasons. First, it is simply not clear that the argument of this essay depends on there being a coherent notion of qualia. Certainly expressions such as 'qualitative character of experience,' and the like, are not required. All that is necessary is to ask whether Sam's pain at t would cease as a result of incapacitating the unused brain region, $R_B$, at t. Therefore, the burden of proof must fall upon the enemy of qualia to show otherwise.

Second, even if it can made out that qualia-based intuitions are at the core of Premise (4), still, one must point to specific ways in which the notion of qualia is confused, and show that one or more of those very confusions are responsible for the intuition behind (4). Dennett asks whether cases of attitude changes toward the same quale can be distinguished from cases in which the same attitude is held toward distinct qualia. He concludes the concept of qualia is not up to the task. But even if that is correct, it is hard to see what relevance that has to this paper (assuming his argument is not taken to demonstrate the nonexistence of qualia). Premise (4) has nothing to do with attitudes toward qualitative experiences. I think it can be safely concluded, therefore, that Dennett's discussion of Chase and Sanborn poses no difficulty for Premise (4). Perhaps some other confusion about qualia is at the root of intuition expressed in (4). But showing that requires an argument.

## 3 Two Related Arguments

In this final section I consider two arguments that are similar, in different respects, to the argument in section 1. My aim is to bring into clearer focus what is distinctive about the argument presented above. First I discuss a recently published argument of Tim Maudlin's. And then I briefly consider the standard, and often discussed, anti-behaviorist "paralysis argument."

### *3.1 Maudlin's Argument*

In 'Computation and Consciousness', Maudlin's (1989) presents an argument against computationalist theories of consciousness similar to the one above. He employs an ingeniously constructed Turing Machine realization called 'Olympia' that runs any program based on any computational theory of consciousness. Though Olympia is too complex to describe fully here, it suffices to say that she is designed to have the following property: With respect to any temporal succession of machine states, $S_1,...S_n$, that instantiates any conscious mental kind, Q, Olympia is constructed so that each in that sequence of states can be causally isolated from all of its conditional effects without any apparent damage to the integrity of Q. Let me clarify.

Suppose Olympia is about to instantiate Q, i.e., run through the sequence of states $S_1,...,S_n$. For Olympia to make the transitions, $S_1$ to $S_2$, $S_2$ to $S_3,...,S_{n-1}$ to $S_n$, the appropriate symbol must be on the tape at each point in the computation. If a symbol other than the one required appears on the tape at any point, Olympia will embark on a new course by instantiating a new sequence of states, and hence Q will not be realized. To instantiate Q by running through $S_1,...,S_n$, therefore, Olympia must be constructed so that if other symbols had been on the tape, she would have gone into the states required by the presence of those symbols; she must be constructed so that the counterfactuals that specify the effects $S_1$, $S_2$, etc. ***might have had*** are true of Olympia. And since Olympia is a realization of a conscious entity, that is likely to require some extensive machinery.

What is significant about Maudlin's construction of Olympia is that the hardware that serves to realize Q ($S_1,...,S_n$) is distinct from all the hardware necessary to realize anything else Olympia might have done, had the symbols on the tape been different. So, because of Olympia's design, when Q is instantiated, activity is required in only a small region of Olympia, and the remainder of her vast machinery can sit absolutely idle. As the reader might now anticipate, what Maudlin argues is that incapacitating the inactive machinery, while Olympia runs through $S_1,...,S_n$, can have no bearing at all on Olympia's phenomenal experience. As Maudlin puts the point, consciousness must supervene on the *physical activity* of a system.[13]

Here is what seems to me to be the main difference between Maudlin's argument and the one above. While both exploit the intuition that unused hardware at t is irrelevant to one's phenomenology at t, Maudlin focusses on hardware that is used to realize effects that any of $S_1,...,S_n$ *might have had* - "counterfactual effects," as it were. The argument in section 1, however, centers on hardware that is used to realize *actual effects* Q is disposed to have at a point *later in time*. A further difference is that the above argument is concerned with the effects *of Q*, whereas Maudlin's focusses on the effects of *constituents of Q* (any of $S_1,...,S_n$) which,

---

[13] It is worth making clear that the intuition behind Premise (4) of the argument of this paper is not that *inactive* brain regions at t are irrelevant to one's phenomenology at t. Rather it is that brain regions at t *not serving to instantiate or realize any mental kinds*, i.e., brain regions *not being used*, are irrelevant. It is entirely accidental (if indeed true) that neural *activity* realizes mental kinds. It might have been (or indeed might be) *structural* properties caused by neural activity that count. I think the same point applys to Maudlin's argument; however he fails to make it, and I do not know whether he would accept it.

presumably, need not themselves involve consciousness.

An advantage of Maudlin's argument is its level of detail and precision, which is displayed through Olympia's elegant construction. But precision often comes at a price, and the price here is that Maudlin's argument lacks the generality of the argument above. It does so in two respects. The first is that Maudlin's argument applies only to computational theories of consciousness, whereas the argument from section 1 applies more generally to functionalist theories. Olympia realizes only programs, but it may be that symbol manipulation is inessential to phenomenal experience. The argument above, on the other hand, makes no reference to symbols.

Second, if computationalism about consciousness turns out to be true, it is plausible that it will be only a weakened version according to which parallel processes are essential to most, if not all, phenomenal kinds. Experiences often seem to have simultaneously instantiated "experiential parts" that evolve through time more or less independently of one another. An experience that consists simultaneously of an auditory experience of a song on the radio, and a visual experience of a walking person would be one example. Now it may be that parallelism is required metaphysically to realize such an experience. If so, however, Olympia cannot run the program. This, however, poses no difficulty for the above argument.[14]

---

[14] Another price of aiming for increased precision by way of more detail, of course, is that there is more that can go wrong. There are two features of Olympia that, on the surface at least, suggest that Maudlin's argument, as it stands, fails entirely. I mention them briefly. Olympia's design makes it impossible for the states constituting Q (i.e., $S_1,...,S_n$) to be *caused* in any of the ways computational theories of consciousness normally would specify. Therefore it is open to the

There is more that could be said about Maudlin's argument and its relation to the argument of this essay. There are some deep, as well as many superficial, similarities and differences between the arguments. A thorough discussion of these, and their implications for consciousness, however, is not possible here. For now I leave these matters as they stand.

### *3.2 The "Paralysis" Argument*

A standard way of arguing against philosophical behaviorism (roughly, the view that mental kinds are dispositions to behave) is to imagine a completely paralyzed person,[15] and claim that, contra behaviorism, there is every reason to suppose such a person could be the subject of mental states, events, etc.[16] This argument, moreover, is equally effective against functionalism, provided the same mental-behavioral connections are specified in the functionalist's definitions. Indeed, it might seem that the paralysis argument can be taken completely "inside the head" and applied to *any* causal connection specified by the functionalist. But if so, the

---

computationalist simply to deny that Olympia realizes Q. Also, Olympia is built so that it is impossible for Q to have different effects depending on what other mental kinds are instantiated: whenever Q is realized, whenever Olympia computes $S_1$ through $S_n$, *precisely the same effect must follow.* But that is likely to be true of few, if any, conscious kinds. Thus one might doubt whether Olympia realizes a conscious entity at all. Maudlin may be able to respond to these worries, perhaps by performing a little surgery on Olympia. But some fix appears to be needed.

[15] For example, a quadriplegic with paralysis of the facial muscles, whose internal organs, however, function properly.

[16] See Shoemaker (1976) for one discussion of this argument.

paralysis argument begins to look remarkably like the argument of this paper. At which point one might reasonably wonder just how the two arguments are related.

The paralysis argument and the one above are similar in the following obvious respects: Both involve the paralysis (incapacitation, destruction, etc.) of tissue that is used to realize actual effects that mental kinds are normally disposed to cause. And both appeal to cases of paralysis to show that such dispositions are inessential to the identities of those mental kinds. In these ways the arguments are alike. However there are three important respects in which they differ. I conclude this essay by briefly pointing them out.

First, in the paralysis argument, *specific* mental-behavioral connections are considered (e.g., between a pain and saying 'ouch'). Based on those considerations, it is then concluded that such connections are inessential, largely because of the existence of actual cases of paralysis where we quite naturally attribute mentality. By way of contrast, the argument from section 1 does not depend on considering cases in which any specific causal relation is severed; and, *a fortiori* not on how we have judged actual cases. On the contrary, it abstracts entirely from these particularities.

Second, with regard to an entity's experience at t, it is crucial to the argument of this essay that one consider the relevance of incapacitating brain regions *that are not in use* at t. That, however, is not so of the paralysis argument. It is irrelevant to the paralysis argument whether a paralytic who is in pain at $t_1$ would say 'ouch' at $t_1$ if he could, or only later at $t_2$. This irrelevance is explained by the first point: Since the paralysis argument aims to show that specific causal

connections are inessential because of the *kinds* of connections they are, no more is needed; in particular, no reference to time is needed, no reference to the fact that incapacitated brain regions are not yet instantiating mental kinds. This point, it seems to me, gets at the most fundamental difference between the two arguments.[17]

A final important difference is this: The paralysis argument is not directed at conscious mental kinds in particular, but applies equally well to kinds that do not involve consciousness, such as nonoccurrent belief, for example. However the opposite seems to be true of the argument of this essay. If one attempts to employ the above argument against a functionalist account of nonoccurrent belief, one will fail.

Consider some nonoccurrent belief of Sam's, one he has not brought to consciousness for some time. Suppose that belief is stored locally. If the neural connections that lead away from the part of Sam's brain that realizes that belief were severed, so that the belief could bring about none of it's normal effects, it seems Sam would cease to have that belief. (I assume the neural connections are not in use when cut.) It would be impossible for Sam to bring the belief to consciousness, the belief could play no role in causing behavior, and so on.

Now even if one thinks it is not entirely clear that Sam would lack the nonoccurrent belief, that is beside the point. For the point is just that arguing that Sam would *have* the belief when the connections are cut, is unpersuasive. That conclusion cannot be reached by way of the argument from section 1, which, unlike

---

[17] On this point, Maudlin's argument and my own coincide. The paralysis argument and my own, however, apply to actual effects of a mental kind, unlike Maudlin's which concerns counterfactual effects.

the paralysis argument, is touching on something distinctive about consciousness.

## REFERENCES


Block, N. (1978). "Troubles With Functionalism," in C. Wade Savage (ed.), *Perception and Cognition, Issues in the Foundations of Psychology, Minnesota Studies in the Philosophy of Science, Vol. 9.* Minneapolis: University of Minnesota Press, 261-325.

Block, N. (ed.) (1980). *Readings in Philosophy of Psychology, Vol. 1.* Cambridge: Harvard University Press, 1980.

Dennett, D. (1978). "Toward a Cognitive Theory of Consciousness," in *Brainstorms.* Cambridge: M.I.T. Press.

Dennett, D. (1987) *The Intentional Stance.* Cambridge: MIT Press.

Dennett, D. (1988). "Quining Qualia," in A. Marcel & E. Bisiach (eds.), *Consciousness in Contemporary Science.* Oxford: Oxford University Press.

Jackson, F. (1982). "Epiphenomenal Qualia," *Philosophical Quarterly* **32.**

Searle, J.R. (1980). "Minds, Brains, and Programs," *The Behavioral and Brain Sciences* **3.**

Lewis, D. (1980) "Mad Pain and Martian Pain," in Block (1980).

Lycan, W.G. (1987). *Consciousness.* Cambridge: MIT Press.

Maudlin, T. (1989). "Computation and Consciousness," *The Journal of Philosophy* **86,** 407-432.

Millikan, R.G. (1984). *Language, Thought, and Other Biological Categories.* Cambridge: MIT Press.

Nagel, T. (1974). "What Is It Like to Be a Bat?" *Philosophical Review* **82.** Reprinted in Block (1980).

Shoemaker, S. (1976). "Embodiment and Behavior," in Amelie Rorty (ed.), *The Identities of Persons.* Berkeley and Los Angeles: The University of California Press. Reprinted in Shoemaker (1984).

Shoemaker, S. (1981a). "The Inverted Spectrum," *The Journal of Philosophy* **74,** 357-381. Reprinted in Shoemaker (1984).

Shoemaker, S. (1981b). "Some Varieties of Functionalism," *Philosophical Topics* **12,** 83-118. Reprinted in Shoemaker (1984).

Shoemaker, S. (1984). *Identity, Cause, and Mind.* Cambridge: Cambridge University Press.

Essay 2

# SOCIAL RELATIONS AND THE INDIVIDUATION OF THOUGHT

## 1 Introduction

Tyler Burge has argued that a necessary condition for an individual's having many of the thoughts he has, instead of others or none at all, is that he bear certain relations to objects (events, properties, states, etc.) in his environment. Such objects include those in the extensions of expressions used to provide the contents of the individual's thoughts, as well as other language users. By way of thought experiment, Burge invites us first to imagine, counterfactually, that the individual lacks the relevant relations, and then to judge that he cannot correctly be attributed many of the thoughts he actually has.[1]

That the natures of many of one's thoughts depend on social relations one bears to language users is an idea Burge has developed since 'Individualism and the Mental'.[2] His main argument for that thesis, however, is found in 'Individualism and the Mental', and rests on his thought experiments involving conceptual error (about arthritis, brisket, contracts, etc.) on the part of the thinker. Of all of Burge's

---

[1] For the thought experiments that are intended to show that social relations to other language users are crucial to the natures of one's thoughts, see Burge (1979, and 1982a). That bearing (causal) relations to natural kinds, artifacts, events, etc., is crucial to which thoughts one has is emphasized in Burge (1982a, 1982b, and 1986c). Finally, for thoughts experiments purporting to show that the contents of low level perceptual states, processes, etc., depend on causal relations that obtain between those states and the world, see Burge (1986a, and 1986b).

[2] Primarily in Burge (1986c, and 1989).

thought experiments, these alone support the thesis that social relations are crucial to the natures of one's thoughts; for in these alone are one's social relations all that is manipulated between the actual and counterfactual situations.

My target in this essay are those very thought experiments in 'Individualism and the Mental'. For Burge to derive his conclusions from the experiments, he must make it plausible that, counterfactually, the individual lacks at least one thought he actually has. I shall try to argue that is not plausible. What will result is not an argument for Individualism - since Burge's other thought experiments will remain untouched by what I say - but an argument that one's social relations are inessential to the individuation of one's thoughts.

### *1.1 Burge's Thought Experiment*

Here, briefly,.is Burge's familiar thought experiment from 'Individualism and the Mental'. (Those initiated in the Burgean Mysteries might skip this paragraph and the two that follow.) The thought experiment has three steps. In the first we imagine an individual, Yolanda, who has many beliefs, occurrent and nonoccurrent, that can be correctly attributed to her with that-clauses containing 'arthritis' in oblique occurrence. She believes that her father has arthritis in his ankle, that arthritis is painful, and so forth. Also, and crucially, she believes falsely that arthritis can be had in the thigh.

In the second step we imagine a counterfactual situation in which Yolanda's physical and phenomenal histories, as well as her nonintentionally described

dispositions, are held constant. The situation is counterfactual in that 'arthritis' is correctly used in the community in a way that encompasses Yolanda's actual misuse. That is, counterfactually 'arthritis' *does* apply to ailments in the thigh (and perhaps elsewhere), in addition to arthritis.

The third step, finally, is an interpretation of the counterfactual situation. We are invited to judge that Yolanda lacks most or all beliefs attributable with 'arthritis' in oblique occurrence. The word 'arthritis' in Yolanda's language community does not mean *arthritis*, and we can suppose no other word in her repertoire does. We might even imagine that no one in the counterfactual situation has ever isolated arthritis for special consideration. Under these circumstances, Burge would submit, it is hard to see how Yolanda could have picked up the notion of arthritis. But if she lacks that notion, she cannot correctly be attributed beliefs with 'arthritis' in oblique occurrence, and so her thoughts in the counterfactual and actual situations differ.

That, then, is the thought experiment. It is important to note that Burge does not take the conclusion of the experiment (viz., that Yolanda's actual and counterfactual thoughts differ) to be entailed by his description of the actual and counterfactual situations, or by anything else he says. While he defends at length the point that Yolanda actually has arthritis thoughts[3] in spite of her conceptual error, he offers little defense of the claim that counterfactually she lacks arthritis thoughts, and maintains only that 'it is plausible, and certainly possible' that she

---

[3] By 'arthritis thoughts' I mean thoughts correctly ascribable with 'arthritis' in oblique occurrence. And *mutatis mutandis* for expressions other than 'arthritis'.

does. That claim is meant to rest primarily, if not entirely, upon our judgments or intuitions about Yolanda's thoughts in the counterfactual situation.[4] Now Burge does provide an *account* from which it follows that Yolanda's thoughts differ in the actual and counterfactual situations - his story that "language community membership" is essential to the natures of one's thoughts. But that is a *result* of the thought experiment: it derives what support it has in virtue of its being an *explanation* of the intuitions the thought experiment generates.

### *1.2 Preview*

All of this, consequently, leaves room for a quite different account of what underwrites the thought experiment, of why we have the intuitions we do; a story, moreover, on which Yolanda's actual and counterfactual thoughts are *the same*. Part of what I shall attempt to do in this essay is to provide such a story. In particular, I shall offer an *alternative explanation* of the thought experiments on which one can accept the intuitions Burge wishes to evoke while denying his conclusions. That, ultimately, will involve sketching an account of propositional attitudes on which the actual and counterfactual thoughts of those in the experiments are the same, and then showing how that is consistent with the intuitions the experiments generate.

The alternative account on its own, however, is not enough to refute Burge;

---

[4] Burge's defense of the first step of the thought experiments to be found, primarily, in section 3 of 'Individualism and the Mental' (Burge, 1979). And his claims to the effect that the conclusion of the thought experiment is meant to rest on it's intuitive plausibility, and not on any particular theory from which it follows, for example, are to be found in Burge (1979, pp. 88-89; and 1982, p. 288).

it still must be shown that it is preferable to Burge's account of the experiments. This I attempt to do by eliciting some intuitions Burge does not consider, and then arguing that his picture cannot accommodate them. The alternative account, on the other hand, explains those intuitions naturally. The argument is thus complete.

The remainder of the paper is structured as follows. In section 2, I develop the alternative account in abstraction from questions about propositional attitudes and attitude attributions in particular. On the basis of general semantic and metaphysical considerations, I show how the kinds of intuitions generated in Burge's thought experiment can result from the attribution of properties that are very different from the sort Burge takes propositional attitudes to be. The general account having been given, I then show in section 3 how the alternative account applies to propositional attitudes specifically. Finally, in section 4, I argue that the alternative explanation ought to be favored over Burge's. I end with some remarks about the social character of thought.

## 2 Developing the Alternative Account

In general terms, here is what occurs in the thought experiment: In the first step, we ascribe a predicate **F**[5] to an object o (we ascribe, e.g., the predicate 'believes that arthritis is painful' to Yolanda). In the second step, we imagine a counterfactual situation in which o's intrinsic properties are held constant, but at

[5] I shall use boldface type rather than quotation marks to name dummy expressions.

least one of o's relations are changed (e.g., Yolanda's relation to her language community). So, whereas in the actual situation o bears a relation, R, to some object p, in the counterfactual situation that is not so (though o may bear R to something else). Finally, in the third step, we judge **Fo** to be false in the counterfactual situation.

It will be useful, for the discussion that follows, to think of the thought experiment as involving an individual (e.g., Burge's reader, my reader, etc.) who, in two separate *contexts*, and with respect to two distinct *circumstances of evaluation*, makes a judgment concerning the truth value of **Fo**. The two "contexts of judgment" are distinguished from each other at least with respect to the *time* at which the judgments are made (the first step of the thought experiment is taken before the third), though below I shall suggest there are other important differences as well. And the two circumstances of evaluation with respect to which **Fo**'s truth value is judged are the actual and counterfactual situations. Accordingly, we can say this of the thought experiment: in the first context, with respect to the first circumstance of evaluation, **Fo** is judged to be true, and in the second context, with respect to the second circumstance, **Fo** is judged false. Now what any account of the thought experiment must provide is an explanation of why this change in judgment occurs across the two contexts and circumstances of evaluation.

So how can it be explained? One way - the route Burge takes - is to say that the predicate **F** expresses a *relational property*, much as 'brother', 'planet', and 'sunburn' do.[6] Specifically, one claims that **Fo** is true in the first circumstance of

---

[6] The point is not that the predicate expresses a relation between an individual and a *proposition*. That relation can be ignored in the rest of this section. The

evaluation (the actual situation), in part, because o bears R to p. Or, equivalently, one says that if o were not to bear R to p, **Fo** would be false. And that, of course, provides a direct explanation of why **Fo** is judged to be false in the second circumstance of evaluation (the counterfactual situation): o does not bear R to p.[7]

So we have one way of explaining the opposing judgments regarding **Fo**'s truth value across the two contexts. But it is not the only way. In the rest of this section, I shall provide an alternative explanation according to which the property expressed by **F**, in both contexts of judgment, *is not a relational property that involves R.*[8] I shall develop a different account of the semantics of **F** which, when taken with the specific characteristics of the two contexts of judgment, and the two circumstances of evaluation, allows for a story of why our judgments of **Fo**'s truth value change that is quite different from Burge's.

### *2.1 Two General Features of the Alternative Account*

---

point, rather, is that the predicate involves some other relation (viz., to a language community) that is not explicitly stated.

[7] To illustrate, consider the predicate 'is a brother'. That predicate applies to Dick Smothers in virtue of his having the same parents as someone else; if that were not so, he would not be a brother. Now if we judge he is not a brother in a world in which his parents have only one child, that is explained by the fact that he lacks the required relations.

[8] I leave open the possibility that **F** might express a relational property that involves a relation *other than R.* In terms of Burge's thought experiment, that means that 'believes that arthritis is painful', e.g., although not expressing a property that involves relations to *language communities*, may involve other relations, such as causal relations to arthritis, pain, etc.

In proposing an alternative account of the intuitions the thought experiment generates, I am, in effect, accepting the correctness of the two judgments of **Fo**'s truth value. I grant it is correct in the first step to judge, with respect to the actual situation, that **Fo** is true; and in the third step, to judge with respect to the counterfactual situation that **Fo** is false. That granted, it would seem any alternative to Burge's explanation must hold that, across the two contexts in which the individual evaluates **Fo**, **Fo** expresses *distinct propositions*. For if the same proposition is true in one circumstance and false in another, and the only relevant difference between the circumstances is **o**'s relations, those relations must somehow be contained in the proposition. If they are, however, Burge's explanation is correct.

The alternative account, therefore, will maintain that **Fo** expresses distinct propositions across the two contexts. In terms of Burge's thought experiment, that will mean, for example, that when the sentence,

(1) Yolanda believes that arthritis is painful

is evaluated in the first and third steps, different propositions are evaluated. But since there is no reason to suppose different *people* are referred to in the two circumstances, the difference must boil down to a difference in the *property* expressed by the predicate 'believes that arthritis is painful'. Accordingly, we can state a general feature of the alternative account thus:

(A) Which property is expressed by **F** can vary across contexts of use

(judgment, utterance, etc.).

I have said that, on the alternative explanation of the thought experiment I wish to develop, **F** does not express, in either context of judgment, a property that involves the relation R. With respect to Burge's experiment, that means the property expressed by 'believes that arthritis is painful' does not involve Yolanda's social relations. I now shall show how that is possible if (A) is true.

A simple example illustrates the point. Imagine two circumstances of evaluation, in the first of which snow is white, and people love snow, and in the second of which snow also is white, but people hate snow. All that differs between the two circumstances, then, are the relations between snow and people. Suppose **F** means *is white*, and **o** refers to snow. **Fo**, therefore, expresses the proposition *that snow is white*. And imagine an individual in some context who correctly judges **Fo** to be true in the first circumstance of evaluation.

Imagine now that there exists another *language* that the individual knows which includes the expressions **F** and **o**. This second language is just like the first, except for the following difference: **F** means *is black*. (**o** still refers to snow.) In the second language, then, **Fo** expresses the proposition *that snow is black*. Suppose now that in a second context, and with respect to the second circumstance of evaluation, the individual is asked to evaluate **Fo**, and for some reason the individual considers the expression **Fo** from the second language. Evaluating the proposition *that snow is black*, then, the individual correctly judges **Fo** to be false.

The first thing to notice is that the difference between the two circumstances, as well as the pattern of judgments across the two contexts, precisely conforms to the general features of the thought experiment: all that differs between the two circumstances are o's relations; and **Fo** is judged true in the first context, and false in the second. Now in judging the truth value of **Fo** across the two contexts, the individual evaluates two distinct propositions. (A), therefore, is satisfied. But it also is true that, in each context, the property expressed by **F** is an *intrinsic property* (viz., being white, being black); more to the point, it is a property that is independent of the relations (between snow and people) that distinguish the two circumstances.[9] In other words, in each context, the property expressed by **F** is had by o in the first circumstance if and only if o has it in the second. If (A) is true, therefore, it is possible for the property expressed by **F** not to involve R.

We can now list a second general feature of the alternative explanation:

(B) Whether o has the property expressed by **F**, in either context, is independent of whether o bears R to any particular object.

Although the above illustration shows how (B) can be true if (A) is, it does not explain why the individual's judgments change across the two contexts. For

---

[9] The distinction I am drawing attention to can be illustrated by letting **F** mean *is heavy* in the first language, and *is light* in the second, instead of the meanings assigned in the text. In that case, the properties attributed to snow in the two contexts are not *intrinsic* properties (since being heavy and being light involve relations to massive bodies), but they still would be independent of people's love or hate for snow. This distinction will be important throughout this section. It is what allows for propositional attitudes to be relational properties, while not involving *social relations* in particular.

even if Fo expresses distinct propositions across the two contexts and circumstances; and even if o actually has the property expressed by F if and only if it has it counterfactually; in order to account for the thought experiment, it still must be explained why we should, unknowingly to ourselves, switch the propositions we consider between the first and third steps of the thought experiment.

That will have to wait until near the end of this section. For now, what needs to be shown how there could be a predicate from a *single language* that satisfies (A) and (B).

### *2.2 Implicative Versus Literal Uses of Relational Predicates*

Consider the sentence,

(2) Sam's car is the same color as a lemon.

An interesting property of sentences like (2) is that they can be used to express two very different sorts of propositions. First, (2) can be used in accordance with its literal meaning to attribute a *relational property* to Sam's car - namely, the property of having a color that is the same as the color of lemons. But second, (2) can be used to attribute to Sam's car an *intrinsic property* - the property of being lemon yellow. To see this second usage more clearly, imagine one wanting to say of some object that it is a certain color, but lacking a word for that color. In such a case, one might point to something else that has the color and say,

'It's the same color as *that*'. And in so doing, one may be not at all interested in saying merely that the two objects have the same color - as if it were unimportant what the color is. On the contrary, the color would be of primary importance, for it would be precisely the property of being *that color* one wishes to attribute to the first object.

Since the first way of using (2) attributes the property corresponding to the literal meaning of the predicate 'is the same color as a lemon', I propose to call it the *literal use*. On the second sort of usage, what is said is something that is *implied by* the proposition that corresponds to the literal meaning of (2), along with a fact about the object, or class of objects, referred to in the predicate. That (a) Sam's car is the same color as a lemon, and that (b) lemons are lemon yellow, implies that (c) Sam's car is lemon yellow. And that implication (c) is what is expressed by the second usage of (2). Accordingly, I call the second usage the *implicative use*.

An important contrast between literal and implicative uses of (2) shows up when one considers the truth value of the expressed proposition across circumstances of evaluation which conform to the general features of the thought experiment. So imagine one circumstance in which both Sam's car and lemons are lemon yellow, and a second counterfactual circumstance in which Sam's car is lemon yellow but lemons are blue. On the *literal use* of (2), since it is the relational property of being the same color as a lemon that is attributed to Sam's car, the proposition expressed by (2) is true in the first circumstance and false in the second. On the *implicative use*, however, where (2) attributes the property of

being lemon yellow to Sam's car, the proposition expressed is true in *both* circumstances. For Sam's car *is* lemon yellow in both.

Notice something else. Take the counterfactual circumstance in which Sam's car is yellow and lemons are blue. And imagine a context in which one is speaking of that circumstance, and using (2) *implicatively*. If one says implicatively, with regard to that circumstance, that Sam's car is the same color as a lemon, what one expresses is the false proposition that Sam's car *is blue*. Not the true proposition that it is yellow. For in such a context of utterance, the expression 'lemon' refers to lemons in the counterfactual circumstance. Across the two contexts of utterance, therefore, implicative uses of (2) express *different propositions*. With *literal* uses of (2), on the other hand, the story is different: in both contexts the *same proposition* is expressed - the proposition that Sam's car has the same color lemons do. Finally, notice that just as the proposition implicatively expressed in the first context (that Sam' car is lemon yellow) is true in both circumstances, so is the property expressed in the second context (that Sam's car is blue) *false in both circumstances*.

With implicative uses of (2), then, we have a sentence which, in two contexts of use, and with respect to two circumstances that conform to the thought experiment, contains a predicate that satisfies *both (A) and (B)*. Different properties are attributed to Sam's car across the two contexts (A); and since, in both contexts, the property expressed by 'is the same color as a lemon' is had by Sam's car in the first circumstance if and only if Sam's car has it in the second, neither of those properties involve relations to lemons (B).

### *2.3 Adding Indexicals*

There is a problem, however, with modelling the alternative account on implicative uses of the predicate 'is the same color as a lemon'. Although in each context, the property implicatively expressed by that predicate is had by Sam's car in the first circumstance if and only if in the second, which property one expresses by that predicate seems to be determined by which circumstance one speaks of. There appears to be no way, for example, to implicatively use (2) to say *with respect to the counterfactual circumstance* that Sam's car is lemon yellow; one seems forced, rather, to express the false proposition that Sam's car is blue.

It would be desirable, however, if the alternative account of propositional attitude-ascribing predicates could be such that not only does Yolanda have all the same propositional attitudes in the actual and counterfactual circumstances, but one can *say what they are when considering her in those circumstances.* And that would most easily be satisfied is if it were possible to use the same attitude-ascribing sentences with respect to both circumstances - if it were possible, for example, to truly *say*,

(1) Yolanda believes that arthritis is painful

not only of Yolanda in the actual situation, but of her in the counterfactual situation as well.

To do that, however, would be both to deny that (1) is true in the counterfactual circumstance - in order to account for the judgments of the thought experiment; and to affirm the truth of (1) in that circumstance - since the true proposition expressed by (1) with respect to the actual circumstance is also true counterfactually. And the only way a sentence can be both true and false with respect to the same circumstance is if that sentence contains (perhaps implicitly) *an indexical component.* 'Yolanda is here', for instance, can be both true and false, depending on which proposition it expresses across contexts of use.

This suggests that the way to solve the difficulty posed by (2) is to include an indexical in the predicate, while still focussing on implicative uses of that predicate. Take, for example, the sentence,

(3) Lucy is shorter than average.

Clearly, 'is shorter than average' contains an indexical element. What must be indexed for (3) to express a proposition is a class of objects from the elements of which a particular average height can be determined. Suppose Lucy is five feet tall. If the class of objects indexed in some context is the class of adult humans, then (3) will express a truth in that context, since the average adult human is taller than five feet. If, in some other context, the class of female jockeys is indexed, (3) may express a falsehood.

Once a class of objects has been fixed in a given context of use, (3) can be used, like (2), either literally or implicatively. Imagine the class of adult humans is

indexed. A literal use of (3) correctly attributes to Lucy the relational property of having a height less than the height of the average adult human. (3) used implicatively, on the other hand, correctly attributes to Lucy an intrinsic property - roughly, the property of having a height below a certain upper bound (five feet, seven inches, let us suppose).[10]

Used either literally or implicatively, (3) can function to attribute different properties to Lucy with respect to the *same* circumstance of evaluation. This is illustrated in the paragraph before last. Different properties also can be attributed to Lucy *across* circumstances of evaluation. Imagine two circumstances, one in which the average height of adult humans is as it is actually, and the second in which the average height of adult humans is three feet. Suppose Lucy is five feet in both circumstances. With both the literal and implicative uses of (3), different propositions can be expressed with respect to the two circumstances merely by indexing to classes of different sorts of objects in each circumstance - e.g., female jockeys in the first and adult humans in the second.

If the *same class* is indexed in both contexts of use, on the literal use of (3), *one* proposition would be expressed across both contexts. If it were the class of adult humans, for example, the proposition expressed would be that Lucy is shorter than the average adult human. That would be true in the first circumstance and false in the second. On *implicative* uses of (3), however, with that same class

---

[10] We might imagine one implicatively using (3) to inform someone that Lucy can get under a particular doorway without ducking. And Lucy would still have the property attributed to her in a world in which the average height of adult humans is three feet.

indexed in both contexts, *different* propositions would be expressed. In the first context, what would be expressed is the true proposition that Lucy's height is less than five feet, seven inches. In the second context, what would be expressed is the *false* proposition that Lucy's height is less than three feet. This last case is very similar to the case of implicative uses of (2).

Let us examine, therefore, whether (3) avoids the difficulties (2) presented. The question we must ask is this: In speaking of the second circumstance, can we implicatively use (3) to express the same proposition we expressed with an implicative use of (3) in speaking of the first circumstance? More specifically, with respect to the circumstance in which the average height of adult humans is three feet, can we implicatively use 'Lucy is shorter than average' to express the proposition that she has a height shorter than five feet, seven inches? If we can, then in speaking of the second circumstance, we can index to humans in the *first* circumstance. But it seems we *can* do that. Consider:

(4) In the second circumstance Lucy *is* shorter than average; *everyone* is.

This seems clearly to contain a reading of (3) of the sort we are after. It says that Lucy (and everyone else) has a height shorter than five feet, seven inches - the average height of adult humans in the first circumstance. Notice also, however, that although it may not be natural for *us* to do so, we certainly can imagine a little man in the second circumstance uttering the following sentence:

(5) In the first circumstance, it's false that Lucy is shorter than average;

*no one* is.

Thus it also seems that the false proposition implicatively expressed with (3) with respect to the second circumstance (that Lucy has a height less than three feet), in addition to being false in the first circumstance, can also be *said to be false* in that circumstance with an implicative use of (3).

The indexical element contained in (3), therefore, seems to have provided much of what we need for the alternative account. It is now possible to give a general explanation of how the judgments of **Fo**'s truth value change across the contexts of judgment in the thought experiment.

### *2.4 Explaining the Pattern of Judgments*

Suppose we describe the following thought experiment. In the first step, we imagine that Lucy is five feet tall, and the heights of other adult humans are much as they actually are. Our task is to evaluate (3).

(3) Lucy is shorter than average.

Since the predicate 'is shorter than average' contains an indexical element, some class of objects must be indexed before (3) can express a proposition. Which class, then, do we index? A natural choice is the class of adult humans (or perhaps the class of adult females) in the first circumstance, and for two reasons. First, in the

description of the first step of the experiment, reference is made to that class. It would be reasonable to infer, therefore, that that is the class of objects we are *meant* to index. Second, it is likely that the class of adult humans serves as a *default index* whenever no comparison class is specified. All the more reason, then, to index that class when it is mentioned. So we do, and then judge (3) to be true.

In the second step of the experiment, a different circumstance is described. Lucy is the same height as she is in the first circumstance (five feet), but the average height of adult humans is three feet. In the third step, finally, our task again is to judge the truth value of (3). Again, therefore, we must settle on an index. Which this time? For the same two reasons the class of adult humans in the first circumstance is indexed in the first step, so in the third step is it natural to index to the class of adult humans *in the second circumstance.* And we do, and subsequently judge (3) to be false.

Our judgments, therefore, change across the two contexts. And that is explained by the fact that the predicate 'is shorter than average' contains an indexical, and we switch indexes between the first and third steps of the experiment. And the reason we switch is that *the description of the second circumstance* makes it overwhelmingly natural to do so.

Now notice that it would be natural to index in the above way, and thus produce the change in judgments, *regardless of whether (3) is used literally or implicatively.* For the process by which an index is settled upon in a given context seems independent of which type of usage is employed. (Default indexings, conversational context, etc. would seem to be all that matters on either usage.)

However, if we suppose (3) is used *implicatively* in the judgments of the thought experiment, then in addition to getting the required pattern of judgments across the two contexts, the properties attributed to Lucy turn out not to involve relations she bears to adult humans. The properties, rather, are intrinsic properties, ones she has in the first circumstance if and only if she has them in the second. Moreover, although it is *natural* to index to the class of adult humans in the circumstance being spoken of, it is not *necessary* to do so. In speaking of Lucy in the second circumstance, one can index to objects in the first, and vice versa, as we have seen.

### *2.5 Predicates With Implicit Relations*

All the main elements of the alternative explanation of the thought experiment are now in place, save one. Each predicate considered thus far has been *explicitly relational in form.* More to the point, the predicates have contained expressions that refer to the relation manipulated across the two circumstances of the thought experiment. However, the predicate 'believes that arthritis is painful' for example, contains no expression (indexical or otherwise) that refers to a relation to a language community. Accordingly, it must be made plausible that there can be predicates that have an *implicit* relational form, while still having the features we require.

Such predicates, however, are easy to come by. Consider the predicate 'is short', for example, and the sentence,

(6) Lucy is short.

'Is short' contains an indexical element, since it can be said of the same person in the same circumstance both that she is short and that she is not short. That requires only that a different class of objects be indexed across the contexts in which (6) is affirmed and denied. Indeed, the predicate 'is short' would appear to function just like the predicate 'is shorter than average', except for the fact that being short probably requires that one be shorter than just shorter than average. If we ignore that difference, all the points made with 'is shorter than average' can be made with 'is short'.

Running the thought experiment with (6) instead of (3), we will judge (6) to be true in the first circumstance, and false in the second. That, however, will be because we have switched indexes from adult humans in the first circumstance to adult humans in the second. And again, though switching in that way is natural, given the way the experiment is described, it is not necessary. We can just as well correctly say of Lucy in the second circumstance that she is short, and of her in the first circumstance that she is tall. Finally, 'is short' can be used implicatively. (I suspect that is its typical (if not only) use.) Thus the property we attribute need not involve a relation between Lucy and objects in her environment. Other predicates of this sort are: 'is heavy', 'is expensive', 'is eccentric', among many others.

The general story of the alternative explanation is now complete. It must now be shown how it applies to the specific case of Burge's thought experiment.

# 3 Explaining Burge's Thought Experiment

## *3.1 Burge's Picture*

It is best to tell the alternative story in contrast to Burge's. Consider, once again, the following sentence:

(1) Yolanda believes that arthritis is painful.

On Burge's view of the nature of propositional attitudes, (1) is true if and only if Yolanda is in a belief state with the content expressed by the clause, 'that arthritis is painful'. A necessary condition for being in such a state is that she have the constituent "notions" expressed by the words 'arthritis', 'is', and 'painful', and that those notions be properly structured with respect to each other.

Now under what conditions will she have or lack those notions? As Burge argues at length, she does not require mastery of them; on the contrary, she can be conceptually mistaken regarding them. Indeed, the thought experiment depends on there being such a conceptual error, and it shows up in the first step where Yolanda believes one can have arthritis in the thigh. What Yolanda requires, therefore, is a certain *minimal competence* with the notions expressed in the that-clause. In virtue of what, however, in the actual situation, for example, is she minimally competent vis a vis the notion of arthritis? Certainly she must contribute something. Her

internal (physical) structure must be such that it produces, under normal conditions, and in the appropriate circumstances, the sorts of bodily motions, brain events, sounds, etc., that are required for minimal competence with the notion of arthritis. After all, not all physical structures can be thinkers.

However that is not enough, according to Burge's interpretation of the thought experiment. For Yolanda's internal structure is the same, actually and counterfactually, but she is competent with the notion of arthritis only in the actual situation. It is here that *language community membership* enters the picture.

The story must go something like this. Associated with each language community, there must be a way of fitting notions, or organized schemes of notions, over individuals' internal structures, or parts of those structures. Typically, of course, this gets done indirectly by way of the individuals' behavioral dispositions; though in principle perhaps the route could be more direct. Since different structures might realize the same scheme of notions, we might say that, associated with each community, there is a function from internal structures into notions, or notional schemes. Call this an *interpretation function*, or *I-function*.[11]

Now one thing the I-function associated with the English community does is assign to Yolanda in the actual world the notion of arthritis, in light of her dispositions to utter 'arthritis' when she does, etc. The I-function associated with the counterfactual community, on the other hand, assigns to *those very same dispositions* the notion expressed by *their word* 'arthritis' - the notion *tharthritis*, let

[11] Though the notion of an *interpretation function* could certainly be sharpened, its vague and undeveloped form will suffice for the purposes of this essay.

us say. Moreover, the counterfactual I-function does not assign the notion of arthritis to Yolanda at all; nor does the actual I-function assign to her the notion of tharthritis.

In the actual situation therefore, Yolanda is minimally competent with the notion of arthritis (but not tharthritis) because the I-function associated with the actual language community *says so*, given her internal structure (and perhaps also her causal relations to objects, etc. in the world). That specific I-function, and that one alone, is relevant to which notions Yolanda actually has, on Burge's view, because the linguistic community associated with that I-function is the one in which Yolanda *is a member*. Similarly, in the counterfactual situation, Yolanda's notions are determined by her internal structure, and the I-function associated with her linguistic community - the counterfactual one.

That is about all that needs to be said here concerning Burge's account of propositional attitudes, as it emerges from 'Individualism and the Mental'. Propositional attitudes turn out to be relational properties that individuals have in virtue of their internal structures, causal relations, and linguistic community affiliations. Yolanda's counterfactual thoughts differ from her actual ones because her language community changes across the actual and counterfactual situations. And our judgments change between the first and third steps of the thought experiment because we are sensitive to those facts.

### *3.2 The Alternative Explanation*

The alternative account of the thought experiment can now be given. Consider again (1).

(1) Yolanda believes that arthritis is painful.

According to the alternative story, the semantics of the predicate 'believes that arthritis is painful' is very much like the semantics of 'is short' *on its implicative use*; and according to the alternative story, 'believes that arthritis is painful' has no function corresponding to *literal* uses of 'is short', if indeed there are such uses.

With implicative uses of 'is short', a class of objects is indexed (e.g., actual adult humans), and a property of that class (e.g., the average height of its members) is used to "fix upon" a property of the individual to which the predicate is ascribed - a property that is independent of any relations the individual bears to elements of the class. The property attributed, roughly, is the property of having a height that lies below an upper bound determined by the heights of the members of the class. Further, in saying of an individual in a given circumstance that he is short, it is not necessary that the indexed objects exist in that circumstance. It was for that reason we could say of Lucy that she is short in a world in which she is tallest.

On the alternative account, when the predicate 'believes that arthritis is painful' is used, there is also a class of objects that is indexed - namely, a community of language users with a set of linguistic practices. As with 'is tall', a property of the indexed class is used to fix on a property of the individual to which

the predicate is ascribed. In the case of attitude-ascribing predicates, the relevant property of the indexed language community is *the I-function associated with it.*

Now what sort of property does the I-function serve to fix? Take a concrete example. Consider a context in which (1) is uttered, and the language community in the *actual* situation is indexed. Let us call the I-function associated with that community the 'I-function$_A$'. Now in uttering (1), the property the I-function$_A$ fixes upon and attributes to Yolanda, roughly, is the property of standing in the belief relation to the content that arthritis is painful *in a manner licenced by the I-function$_A$*. Put another way, what is said of Yolanda is that her internal structure, and perhaps also her causal relations to objects in the world, are of a certain kind - namely, a kind that meets the minimal requirements, *according to the I-function$_A$*, for believing arthritis is painful. And that is true: in spite of Yolanda's misconception, her utterances of 'arthritis', e.g., count as expressions of the notion of arthritis, according to the I-function$_A$.

Now notice that this property that is attributed to Yolanda is one she has *independently of which language community she is a member of.* Her internal structure and causal relations meet the minimal requirements for believing arthritis is painful, according to the I-function$_A$, *in both the actual and counterfactual situations.* Structurally, she is the same in both situations, and she also is appropriately causally related in both situations to arthritis, pain, and so on. The property, therefore, does not depend on relations she bears to other speakers.

In a context in which (1) is uttered, and the *counterfactual community* is indexed, a different property is attributed to Yolanda. It is a property fixed by the

I-function associated with the counterfactual community (the I-function$_{CF}$), and it is a property Yolanda *lacks*, since she does not meet the requirements of the I-function$_{CF}$ for having the notion of arthritis. And the reason she does not, we can imagine, is related to the fact that her 'arthritis' utterances get interpreted by the I-function$_{CF}$ as expressions of *tharthritis thoughts*, and quite different dispositions, ones which Yolanda lacks, are needed to have thoughts about arthritis. Now this property too is one that is independent of social relations. And Yolanda lacks it in both the actual and counterfactual situations.

Finally, as in the case of 'is tall', the class of objects indexed need not exist in the circumstances being spoken of. Thus, indexing to the actual language community, it can truly be said of Yolanda *in both the actual and counterfactual situations* that she believes that arthritis is painful. And indexing to the counterfactual community, utterances of (1), with respect to each situation, will be false. On the alternative account, therefore, Yolanda's actual and counterfactual thought *are the same*. Both actually and counterfactually, she has arthritis thoughts and lacks tharthritis thoughts where our attributions are indexed to the actual community; and she has tharthritis thoughts and lacks arthritis thoughts where they are indexed to the counterfactual community.

It can now be said how the alternative account explains Burge's thought experiment. In the first step of the experiment, according to the alternative view, when we attribute arthritis thoughts to Yolanda in the actual situation, our attributions are indexed to the actual community. The reason the actual community is indexed is this: We are told that Yolanda speaks English, and that she lives

among English speakers; so it is natural to interpret her with the I-function associated with the English language community. This "naturalness" can be cashed out in the following way: One can say that the language community of which one is a member is a *default index*, much as the class of adult humans is a default index when attributions of shortness are at issue.

The second step of the experiment describes the counterfactual situation, and in the third step, we are invited to intuit that Yolanda lacks arthritis thoughts. On Burge's view, the reason we have that intuition is that her linguistic affiliation in the counterfactual situation, coupled with her internal structure (dispositions, etc.), excludes her from having arthritis thoughts. And that is because interpretation must be carried out by means of the I-function associated with the community in which the individual being interpreted is a member.

On the alternative view, however, the reason we have the intuition that Yolanda lacks arthritis thoughts is that *we have switched indexes from the actual community to the counterfactual community*. But why has the switch been made? Well, we are told that in the counterfactual situation there is a different language community; and we are told Yolanda speaks that language, and is a member of that community. Since the default indexing is to one's own language community, we switch indexes. After all, interpreting Yolanda in terms of the I-function$_{CF}$ will best enable us to explain and predict her interactions with the people in the counterfactual situation among whom she lives. (That, of course, is why the default index is what it is.) For the I-function$_{CF}$ is the unique I-function associated with that community of speakers. So we make the switch. But crucially, according to

the alternative account, we are not *required* to index the counterfactual community, just as we are not required to think of Lucy as tall in the world in which she is tallest. Because it is *natural* to switch indexes, the intuitions are explained. But since it is not *necessary* (in order to be correct) that the switch be made, Burge's conclusion that Yolanda's actual and counterfactual thoughts differ is rejected.

An implication of the alternative view worth stating explicitly is that an individual is having not just *one thought* whenever he or she utters something, but *an indefinite number of thoughts.* For instance, When Yolanda, in either the actual or counterfactual situation, says 'Arthritis is painful', she is having the thought that arthritis is painful, the thought that tharthritis is painful, and an indefinite number of others, where what are indexed are various language communities. This implication might strike the reader as counterintuitive at best. However, it really should be no cause for alarm; just as one ought not to be concerned over the fact that Larry Bird is not just tall, but also extremely tall, sort of tall, short, very short, etc. Once one settles on an indexed comparison class for whatever purposes one has, Bird gets assigned the degree of tallness determined by the meaning of 'is tall' and the comparison class, *and no other degree of tallness.* And so long as one sticks with that class, none of the indefinite number of other degrees of tallness Bird has need enter one's mind. Similarly, once a particular language community is indexed for whatever purposes one has, Yolanda gets assigned a range of notions and thoughts; and other notions which she might be assigned by other I-functions associated with other communities, she is correctly said to lack. And for the entire duration that that I-function remains in play, other thoughts Yolanda has on other indexings *can*

*be kept wholly out of thought and out of mind* - as they should.

So where do matters now stand? We have Burge's explanation of the thought experiment, and the alternative explanation. On the former explanation, Yolanda's actual and counterfactual thoughts differ; on the latter they are the same. Are there reasons for preferring one account over the other? I believe so. In the following section I shall say what they are.

## 4 Evaluating the Options

Burge's conclusion that the natures of many of one's thoughts depends on one's linguistic affiliation is based upon intuitions generated in a single type of thought experiment. Since the alternative view also explains why we have those intuitions, what is needed to decide among the two views are other considerations of some kind. In this section, therefore, I shall offer considerations I think strongly favor the alternative view. Specifically, I shall attempt to elicit intuitions of the following sort: that it is perfectly appropriate, under many circumstances, to attribute thoughts to an individual by means of an I-function that is associated with a language community other than the individual's own. On the assumption that those intuitions are genuine, I shall then argue that while they are easily accounted for by the alternative view, it is unclear how Burge can accommodate them on his current view.

### *4.1 Other Intuitions*

I begin with a minor variation on Burge's thought experiment, one he himself considers briefly in 'Individualism and the Mental'. Burge calls it the "reversed version" of the thought experiment, and the idea is to keep the actual and counterfactual language communities as they are, but have the individual's conceptual error show up in the counterfactual situation instead of the actual one.[1] So we can imagine that Yolanda, for example, uses the word 'arthritis' in the actual situation in accordance with proper English usage. We can suppose she applies it only to inflammation of the joints, and not ailments of the thigh, and so on. But, then, in the counterfactual situation, Yolanda's use of 'arthritis' will constitute a misconception with respect to the standards of her language community. For counterfactually 'arthritis' *does* apply to rheumatoid ailments outside of the joints, and Yolanda believes otherwise. The upshot, Burge maintains, is the same for this reversed experiment as for the standard one: Yolanda lacks arthritis thoughts in the counterfactual situation, and so her actual and counterfactual thoughts differ.

Burge offers two reasons why he chose to emphasize the standard version of the thought experiment rather than the reversed version. My interest here lies with

[1] See Burge (1979, p. 84).

his second.[2] He writes,

> A secondary reason for not beginning with this "reversed" version of the thought experiment is that I find it doubtful whether the thought experiment always works in symmetric fashion. There may be special intuitive problems in certain cases - perhaps, for example, cases involving perceptual natural kinds. We may give special interpretations to individuals' misconceptions in imagined foreign communities, when those misconceptions seem to match our conceptions. In other words, there may be some systematic intuitive bias in favor of at least certain of our notions for purposes of interpreting the misconceptions of imagined foreigners. (I&M, p. 84.)

In the above quotation, Burge puts his finger on precisely the sort of intuition to which I want to draw attention in this section: the kind that involves the application of *our conceptions* (i.e., the I-function associated with our language community) to individuals who are members of other communities.

By 'cases involving perceptual natural kinds', I assume Burge means thought experiments that involve notions of color (red, blue, green), notions of taste (sweet, sour, bitter), and so forth. The following thought experiment, perhaps, illustrates what Burge has in mind. Suppose in the actual situation that 'red' is used

---

[2] The first reason is that both versions of the thought experiment depend on finding a misconception in Yolanda's understanding; but our intuitions are stronger, and more reliable, concerning the status of misconceptions in *our own* language community. I think the idea is supposed to be that since we speak English, we are in a position to judge that in the actual situation Yolanda has arthritis beliefs in spite of her misconception. Since we are not speakers of "counterfactual English", however, we cannot reliably judge that Yolanda lacks tharthritis thoughts (in the reversed experiment) in spite of her misconception.

according to normal English usage. In the counterfactual situation suppose it applies to the red part of the spectrum, and also to some of the orange part. Next, imagine a fellow, Zachary, whose dispositions are such that he applies 'red' only to red objects. In the thought experiment based on these details, then, Zachary incompletely understands his notion in the counterfactual situation, and his understanding is correct in the actual situation.

Now here is Burge's worry. In the first step of the thought experiment, we properly attribute red thoughts to Zachary in the actual situation. Then, if all goes as it should in the second and third steps, we should have the intuition that he lacks red thoughts when we switch to the counterfactual situation. Intuitively, however, it would appear that he has red thoughts counterfactually. When he says 'I love everything red', for instance, it seems the content of his thought includes the notion of *red*; it seems he is expressing the thought that he loves everything red. He may *hate* orange. Now insofar as it seems that way to us, according to Burge, we have given a special interpretation to Zachary's misconception in the counterfactual situation, because it matches our conception. We have been "systematically biased" in favor of our own notions.[3]

I shall consider Burge's "bias" account of these intuitions below. For now I want to suggest that such intuitions are generated in cases other than those that involve perceptual natural kinds. Indeed, they seem to arise in the reversed experiment *across the board.* Take the reversed version of the arthritis thought experiment, for example, and consider Yolanda's thoughts in the counterfactual

---

[3] See his discussion in Burge (1979, p. 84.).

situation. (Recall that Yolanda applies 'arthritis' only to inflammation of the joints.) Does she have arthritis thoughts? When she says 'arthritis is painful', does she express the thought that arthritis is painful? It seems to me far from obvious that the answers should be 'no'. Granted, the counterfactual community applies 'arthritis' to rheumatoid ailments other than arthritis. But Yolanda does not. Her word applies just to arthritis; whatever she normally calls 'arthritis' *is arthritis*.

In the reverse experiment, therefore, it seems intuitively correct to think of the contents of many of Yolanda's thoughts in the counterfactual situation as including the notion of arthritis - particularly, those thoughts she would express with utterances that contain 'arthritis'. That, of course, is not to rule out the possibility that it might also be correct to say she expresses *tharthritis* thoughts by those utterances; it is just to say that it is easy, and natural to conceive of Yolanda as having arthritis thoughts. Consequently, it would appear that this type of result extends beyond experiments involving perceptual natural kinds, and instead is a general feature of the reversed thought experiment.[4] For all that seems required is that the individual's dispositions in the counterfactual situation not constitute any conceptual errors relative to the I-function associated with the actual language community. And that is guaranteed by the way the reversed experiment is set up.

So we have some intuitions that favor interpreting individuals' thoughts in terms of I-functions that are associated with foreign language communities - '*foreign i-functions*', we might call them. And these are generated when the individual has

---

[4] This is not to deny, of course, that the intuitions might be strongest in thought experiments involving perceptual natural kinds.

no misconceptions relative to those I-functions. But what of cases in which the individual *has* misconceptions relative to the foreign I-function? Is it still possible to have the intuition that the foreign I-function can correctly be applied? Where Yolanda applies 'arthritis' to ailments of the thigh in the standard version of the thought experiment, for example, is it possible to get the intuition that she has arthritis thoughts counterfactually? I shall now attempt to demonstrate how such intuitions can be had.

I begin with an example in which Yolanda's dispositions are different from how they are in the standard experiment, and are instead like those she has in the reversed experiment, except for one difference: while she applies 'arthritis' only to ailments in the joints, she is disposed to restrict her applications to cases of inflammation of the joints *in the hands*. We can imagine she has heard the word applied only to such cases, and has inferred incorrectly that the disease is, specifically, a disease of the hands. She says things like, 'My mother's arthritis is acting up again', 'I've heard Bufferin eases minor arthritis pain', and so on. The actual and counterfactual community's usages of 'arthritis' are the same in the earlier examples. So Yolanda misconceives the notion she expresses by 'arthritis' relative to the I-functions of *both* the actual and counterfactual communities.

Now on Burge's picture, Yolanda has arthritis thoughts in the actual situation, but not tharthritis thoughts, and tharthritis thoughts but not arthritis thoughts in the counterfactual situation. Is it intuitively obvious, however, or even compelling, that she lacks arthritis thoughts counterfactually? If we imagine Yolanda worrying over the swollen joints (i.e., *arthritis*) in her mother's fingers,

while being disposed not to apply 'arthritis' to any ailment not located in joints, it would seem she can be conceived quite comfortably to have arthritis thoughts in the counterfactual situation, so long as we are willing to grant them to her actually; that is to say, so long as her restricted application of 'arthritis' to inflamed joints in the hands does not prohibit her from having arthritis thoughts in the actual situation.

If that is so, however, then we have a situation in which it is intuitively correct to interpret Yolanda's thoughts using a foreign I-function, *even though she misconceives some of her notions with respect to that I-function.* These intuitions are different, therefore, from those elicited in the reversed experiment. Moreover, Yolanda's misconception relative to the I-function$_A$ is no worse, in any obvious sense, than her misconception in the standard experiment (where she is disposed to apply 'arthritis' to ailments in the thigh). The current example differs from the standard experiment, however, in that Yolanda has misconceptions relative to the I-function$_{CF}$ *and* the I-function$_A$. Clearly, that is doing some of the work of generating the intuitions in the current example. And that is fine; for my point here is just to show that misconceptions relative to foreign I-functions do not, in themselves, suffice to block intuitions that the foreign I-functions can be appropriately applied.

That having been said, however, it would seem possible to get the same sorts of intuitions even when, as in the standard experiment, Yolanda has *no misconceptions* relation to the I-function$_{CF}$. Suppose Yolanda's experience is such that she has heard (or seen) the word 'arthritis' applied only to inflamed joints in various parts of the body - in the hands, ankles, hips, elbows, etc. And suppose

those are the sorts of cases that come to mind whenever she hears the word 'arthritis' (or reads it, thinks it, etc.). And now, imagine that in spite of her thoughts and experiences being that way, she nevertheless is disposed, *unknowingly to herself*, to apply 'arthritis' to rheumatoid ailments in the thigh. Unconsciously, she has that disposition, but no situation has ever yet arisen that would cause the disposition to "play itself out", as it were.

Let us focus now on what Yolanda has in mind when she asks her mother, "Is your arthritis bothering you today?", while attending to the inflamed joints in her mother's hand. Or when she says, thinking of various examples of inflamed knees, ankles, elbows, etc., "I've heard Bufferin eases minor arthritis pain." Is it clear in this case that it is wrong to conceive of Yolanda as having arthritis thoughts? It strikes me that it is not. Even though she harbors no misconceptions relative to the I-function$_{CF}$, but does relative to the I-function$_{A}$, given the sorts of examples she consciously associates with the word 'arthritis', it seems not difficult at all to conceive of her as a *bona fide* arthritis thinker. But if that is so, we have a case very much like the standard experiment, except for the fact that Yolanda's disposition to apply arthritis to the thigh has been kept "dormant", and irrelevant to the situations in which we are considering her thoughts.

Thus far in this section I have been attempting to elicit intuitions that differ from those Burge focusses on in 'Individualism and the Mental'. More precisely, my concern has been with generating intuitions that individuals can correctly be attributed thoughts by way of foreign I-functions. In the first case that I considered - the case of the reversed thought experiment - the individual had no misconception

relative to the foreign I-function; in the next two, such misconceptions were involved. Although some of the intuitions associated with these cases are more robust than others, no doubt, they would all appear to be genuine. Consequently they must be accounted for in some way. And though it is possible that still other intuitions could be generated, there is now a large enough collection for the purposes of this essay. I now wish to consider what the alternative view and Burge's view might have to say of them.

### *4.2 The Alternative View*

With regard to the alternative view, there is really not much that needs to be said. For the view seems tailor made to accommodate the range of intuitions facing us. In all of the above examples, according to the alternative account, Yolanda has the same thoughts counterfactually and actually. In both the actual and counterfactual situations, she can correctly be attributed thoughts by way of the I-function$_A$, and by way of the I-function$_{CF}$. So if we have intuitions that, counterfactually, Yolanda *has* arthritis thoughts (via the I-function$_A$), *we are right* according to the alternative view. And the reasons it might be natural to attribute those thoughts to her in the above examples, rather than (or in addition to) tharthritis thoughts, might be that she has no misconceptions relative to the I-function$_A$, and so it is easier to think of her in those terms; or that those misconceptions she does have do not interfere with the manner in which we want to think of her, given the context in which we are attributing to her thoughts; and so

forth. In short, therefore, the alternative view accords perfectly well both with the intuitions Burge elicits in the standard experiment, and with the intuitions generated by the above examples.

### *4.3 Burge's View*

The situation is rather different when it comes to Burge's view. On his account, as we have seen, an individual can be attributed thoughts only by way of the I-function associated with the linguistic community of which the individual is a member. On Burge's view, therefore, it is simply false that Yolanda has arthritis thoughts in *any* of the counterfactual situations I have described. Assuming the intuitions are genuine, therefore, Burge must end up saying something like this: the intuitions are illusions; we are being misled into thinking Yolanda can correctly be conceived as having arthritis thoughts counterfactually; we are being systematically biased to adopt our own conceptions by the particular features of the above examples; etc.

We have seen that this is precisely the line Burge takes in the case of the intuitions generated in the reversed experiment, and he must take the same line in regard to the intuitions associated with the other examples. I shall consider this line of response presently. First, however, I wish to discuss a response I believe Burge would not make, but which one might think he should make, or at least could make.

Perhaps it could be argued that of all the intuitions considered in this section that run contrary to Burge's view, the strongest are those that are connected with the reversed experiment. If one is of that mind, and particularly if one also thinks the other intuitions are considerably weaker, one might suppose Burge should adopt the following view: that individuals can be attributed thoughts either with I-functions associated with their language communities, or with I-functions relative to which they have no misconceptions. This view would require that Burge abandon the reversed experiment as a means for establishing his thesis that Yolanda's actual and counterfactual thoughts differ, since Yolanda would *have* arthritis thoughts in the counterfactual situation (as well as tharthritis thoughts). However, the standard experiment would still show that Yolanda's actual and counterfactual thoughts differ; for she would have arthritis thoughts actually, and lack them counterfactually. (And she would have tharthritis thoughts both actually and counterfactually.) Accordingly, this approach might seem to offer a way to Burge of accepting the legitimacy of the contrary intuitions in the reverse experiment, while at the same time retaining his general conclusion. And the idea then would be to either deny having the other (weaker) intuitions associated with the other examples, or to handle them in some other fashion.

My reasons for thinking Burge would not take this line are twofold. First, given the opportunity, he simply *does not* take it; he offers his "bias" response instead. And second, it is rather far removed from both the spirit and letter of Burge's view to allow that thoughts can properly be attributed relative to I-functions associated with foreign language communities - even if there are no misconceptions

relative to those I-functions.

Regardless of what Burge did or would do, however, the view under consideration has a serious difficulty: in certain cases it entails that it must be said of the same individual, and at the same time, that that individual believes both P and not-P, where intuitively the individual believes no such thing.

The argument for this depends on a Burge-like thought experiment in which the individual misconceives not just one, but two of his or her notions with respect to some I-function.[5] Suppose, in the actual situation, that Yolanda's dispositions concerning the words 'arthritis' and 'rheumatism' accord with proper English usage. And imagine that in the counterfactual community, 'rheumatism' refers to arthritis, and 'arthritis' to some specific kind of arthritis - gout, say. With respect to the I-function$_{CF}$, therefore, Yolanda misconceives the notions she expresses with both words 'rheumatism' and 'arthritis' (i.e., the notions *arthritis* and *gout*, respectively).

Now consider Yolanda's thoughts in the counterfactual situation. Suppose she says, 'One can have rheumatism in the thigh, but not arthritis'. According to the I-function$_{CF}$, that utterance expresses the thought that one can have arthritis in the thigh, but not gout. Now on the view we are considering, one can be attributed thoughts relative to any I-function with respect to which one has no misconceptions. And Yolanda has no misconceptions with respect to the I-function$_{A}$. So by way of the I-function$_{A}$ she gets the thought that you can have rheumatism in the thigh but not arthritis. But now we have Yolanda believing both that one can have arthritis

[5] I am indebted to Bob Stalnaker for the type of example on which this argument is based.

in the thigh, and that one cannot have arthritis in the thigh. But intuitively, she has no such contradictory thoughts: there appears to be no class of objects, events, etc., to which she both ascribes, and withholds ascribing, any property P (e.g., *rheumatism, arthritis, gout*, or any other).

It would seem that the only way to avoid attributing contradictory beliefs to Yolanda, on the view we are considering, is to maintain that the sense in which she believes that P is different from the sense in which she believes that not-P. However that brings one very close, if not all the way, to the alternative view. For it amounts to saying that in the context in which the attribution is made of Yolanda, 'believes that P' refers to one property in its first occurrence, and another in its second. But that means there must be some indexical element in the context of utterance that determines which property is expressed by the that-clause in its various occurrences; and that is the heart of the alternative view.

Taking this route, therefore, forces one in the direction of the alternative view. Consequently, it is a route Burge ought not to take if he wants to maintain the essential character of his view. Instead he must stick with the route he follows in 'Individualism and the Mental', and attempt to explain away all the contrary intuitions we have considered in this section as illusory. He must say, as he does with respect to the contrary intuitions generated in the reversed experiment, that there is "some systematic intuitive bias in favor of...our notions for purposes of interpreting the [thoughts] of imagined foreigners".

So what of this line of response - in effect, that we are biased to incorrectly apply the I-function$_A$ in interpreting the thoughts individuals in the counterfactual

situation? Well in the presence of the alternative view that maintains that foreign I-functions *can* appropriately be applied, the response is no response at all. It merely begs the question. We need *reasons* to suppose that the intuitions elicited in this section involve some systematic bias that tricks us into thinking what we think, instead of supposing that the intuitions are *accurate*, as the alternative view would have it. To simply state that the intuitions are illusory, and that we are being systematically misled, is just to say that Burge's view is correct, and that there is something funny going on with the troublesome intuitions. But obviously that will not do; for the intuitions make no trouble for the alternative view.

The bias response in itself, therefore, is not enough. Burge needs something more. However it is hard to see what that something could be, other than an independent argument that his view is correct; that is to say, an argument the conclusion of which is that foreign I-functions cannot correctly be applied. Now notice that if he were in possession of such an argument, he would not require his thought experiments. Since he takes his conclusions to rest upon the judgments elicited in his thought experiments, however, we can infer that he does not take himself to be in possession of the sort of independent argument we now see he requires. None of this, of course, is to say that such an argument could not be had. It is just to say that he is in need of one now, and so he currently lacks an adequate response to the challenges posed by the intuitions that run contrary to his view.

### *4.4 Conclusion*

Since the alternative view *does* accommodate the intuitions, unlike Burge's view, it follows that the alternative view is to be preferred over Burge's. Therefore there is no reason to suppose individuals have the thoughts they do in virtue of being related to speakers in their environment. Thought is not social in the sense in which Burge imagines.

There is a sense however in which thought does retain a social component. According to the alternative view, language communities are indexed in all attributions of thought. In attributing thoughts, therefore, one always appeals to the "principles of attribution" (I-function) of *some* possible language community or other, and the resulting interpretation is essentially connected to those principles. In a sense, then, thoughts are individuated nonindividualistically: thinkers are not considered in vacuums, but rather as they appear under the "conceptual grid" of some possible public language or other. How thoughts get carved up depends upon which grid is used. In this light the alternative view looks not so different from Burge's. However one's linguistic affiliation, one's *social relations*, have no bearing at all, according to the alternative view, on which grids are applicable, and on which thoughts one has. Physical duplicates that are causally related to the same stuff in the world have all the same thoughts. In this respect the views differ.

## REFERENCES


Burge, T. (1979). "Individualism and the Mental," in P. French, T. Euhling, & H. Wettstein (eds.), *Studies in the Philosophy of Mind*, Vol. 10, *Midwest Studies in Philosophy*. Minneapolis: University of Minnesota Press.

Burge, T. (1982a). "Two Thought Experiments Reviewed," *Notre Dame Journal of Formal Logic* **23**.

Burge, T. (1982b). "Other Bodies," in A. Woodfield (ed.), *Thought and Object*. Oxford: Oxford University Press.

Burge, T. (1986a). "Individualism and Psychology," *Philosophical Review* **95**.

Burge, T. (1986b). "Cartesian Error and the Objectivity of Perception," in P. Pettit & J. McDowell (eds.), *Subject, Thought, and Context*. Oxford: Oxford University Press.

Burge, T. (1986c). "Intellectual Norms and Foundations of Mind," *The Journal of Philosophy* **83**.

Burge, T. (1989). "Wherein is Language Social?" in A. George (ed.), *Reflections on Chomsky*. Oxford: Basil Blackwell.

Essay 3

# FODOR AND PYLYSHYN ON CONNECTIONISM

In a recent essay, Jerry Fodor and Zenon Pylyshyn (1988) offer the following argument against connectionism:

(1) The Language of Thought (LOT) is true at the level of cognitive architecture

(2) LOT and connectionism are incompatible

---

(3) Connectionism is false at the level of cognitive architecture

Fodor and Pylyshyn's (F&P's) arguments for (1) and (2) are found in the third and second sections of their paper, respectively.[1] Though the above argument clearly is

---

[1] F&P's argument is in fact slightly different from the one I have displayed. Though the form of their argument is the same, their two premises are not. Specifically, in place of the references to LOT in (1) and (2), F&P refer to what they call '*Classical architecture*', which involves both LOT *and* structure sensitive operations. (See their definition of 'Classical architecture' on pp. 12-13, and their statement of how their main argument is going to go on pp. 6-7.) The defense of their first premise, therefore, must also constitute a defense of (1). And though they could defend the claim that Classical and connectionist architectures are incompatible without defending (2), they do not take that route: they argue that connectionism is incompatible both with LOT and structure sensitive operations.

My reasons for focussing on LOT rather than on Classical architecture are twofold. First, since the most important aspects of F&P's discussion of Classical architecture focus on LOT, their main arguments and my criticisms of them can be more clearly and simply presented if I limit my discussion in this paper to LOT. Second, in the appendix to his (1987) book *Psychosemantics*, Fodor argues for (1), and indirectly for (2). Construing the argument as I have, therefore, draws attention

valid, F&P's defenses of (1) and (2), I shall argue, fail to stand up. In particular, their argument for (1) involves an appeal to *inference to the best explanation*, but requires a defense of additional premises they do not supply; and their argument for (2) turns out to be question-begging. (I shall argue also that (2) is false.) Although F&P *do* present connectionism with a serious empirical challenge - one LOT seems able to meet - and in so doing "gain points" for LOT, nothing so strong as (1) or (3) follows. Accordingly, neither has LOT been secured a place at the level of cognitive architecture, nor has connectionism been denied one.

It should be emphasized at the outset that my aim in this essay to criticize only certain of F&P's strongest claims against connectionism and in favor of LOT - viz., that the latter is true and the former false of psychology - as well as their claim that LOT and connectionism are incompatible. Consequently I shall have little to say of the many weaker, though interesting and important points their essay contains. Those points, which are largely empirical in nature, will undoubtedly continue to be discussed in the psychological and AI literatures for some time. My purpose here merely is to clear away some of F&P's more extravagant claims (as I see them), so that the genuine empirical issues can be viewed in a more realistic light.

I proceed as follows. In the first section, I say a few words about cognitive architecture, LOT, and connectionism, so that F&P's argument can be properly understood, and to provide some necessary stage setting for my criticisms of it. That done, I take up F&P's arguments for premises (1) and (2), respectively, in

---

to the close relation between F&P's and Fodor's own arguments.

sections two and three.

## 1 Architectures and F&P's Argument

### *1.1 Cognitive Architecture, Lot, Connectionism*

Cognitive science treats minds as computational devices. In characterizing a computational device *qua* computer, one specifies its *functional architecture* - roughly, the device's *fixed resources*, functionally described. Thus one points to the device's primitive operations, its representational capacities, the features of its memory, and so on. Such a characterization sets out the constraints within which all computation must occur, and in that sense precisely determines what the machine can do.

Insofar as the mind is treated as a computer, its functional architecture too must be characterized. The *cognitive architecture*, as it is typically called, is the functional architecture of the mind, and uncovering it, along with the algorithms that get executed in it, is the primary aim of cognitive science.[2] Connectionism and LOT, then, are proposals about what the cognitive architecture is. Or, more accurately, they are proposals about certain *properties* of the cognitive architecture, and thus are better understood as cognitive architectural *frameworks*, where considerable detail still needs to be fleshed out.

---

[2] This familiar story is spelled out in greater detail in works such as Pylyshyn (1984) and Haugeland (1985), as is the discussion below on *levels of architecture*.

For a functional architecture to be an instance of LOT, according to F&P, it is necessary and sufficient that the representations operated on have a *combinatorial syntax and semantics*, which amounts to satisfying the following three conditions:

> (a) there is a distinction between structurally atomic and structurally molecular representations; (b) structurally molecular representations have syntactic constituents that are themselves either structurally molecular or are structurally atomic; and (c) the semantic content of a (molecular) representation is a function of the semantic contents of its syntactic parts, together with its constituent structure. (p. 12)

Now being an instance of LOT undoubtedly provides a functional architecture with powerful computational capacities it would otherwise lack. However, it is a property that indefinitely many possible functional architectures realize. Most competing architectures in AI, after all, instantiate LOT. It should be clear, then, that LOT provides only a framework for modelling cognitive architecture, and at most captures only a very general property of it.

Connectionism also provides an architectural framework for the study of cognition, rather than a full blown theory. It offers a more detailed account than LOT, however, being closer in its degree of detail to *instances* of LOT like production systems for example, than to LOT itself. Because connectionism is an evolving research program, however, rather than a notion that has been defined on the basis of a small set of properties, like LOT; and also because there are not a few conceptual difficulties associated with our current understanding of the nature of

connectionist computation[3]; it would seem inappropriate at this point to attempt to define 'connectionism', or provide necessary and sufficient conditions on connectionist architectures. Accordingly, I shall simply list the basic architectural features commonly mentioned in general accounts of connectionism. (These probably come close to constituting a set of sufficient conditions on connectionist architectures, though I shall not press the point.)

Models in the connectionist framework typically consist of a set of simple processing units that take on various activation values, and a pattern of connectivity among the units. In the general case, individual units receive input from, and send output to, several other units. Each connection between pairs of units has associated with it a "strength" or "weight" that determines the effect one unit has on another. At any given time, a unit's activation value is a function of its previous level of activity, as well as the degree of activity (excitatory or inhibitory) currently being received along its input connections. Finally, what does the representing in connectionist models are either activated individual units - local representations - or patterns of activity across multiple units - distributed representations.[4]

---

[3] For example, the standard conception of computation in terms of *symbol manipulation*, which involves both symbols and distinct operations on them, does not naturally apply to connectionist machines. Certain connectionists (e.g., Smolensky, 1988) even deny that symbol manipulation occurs in connectionist machines at all. Getting clear on these matters, it seems to me, is an important philosophical task. Some interesting work along these lines can be found in Cummins (forthcoming).

[4] There is some question whether to count activity in so called "hidden units" as representational, and also how to determine what the representational content is, if it is so counted. Moreover, the *weights of connections* are also often said to represent some of the system's knowledge, memories, etc., in addition to unit activity. These are just a few of the murky issues I alluded to in the previous paragraph. I shall ignore them in this essay.

### *1.2 Understanding F&P's Argument*

Consider now the premises and conclusion of F&P's argument, beginning with Premise (2).

(2) LOT and connectionism are incompatible

Since LOT and connectionism are frameworks for, or properties of, functional architectures, (2) amounts to the claim that LOT and connectionism cannot both be instantiated in the same functional architecture. Put slightly differently, no connectionist model can be an instance of LOT (have the property of a combinatorial syntax and semantics), nor can any LOT model be an instance of connectionism (have whatever properties suffice for that: roughly, the sorts described above).

This notion of *instantiation* (where, e.g., a connectionist model instantiates LOT) must be distinguished from a quite different notion with which it can easily be confused - the notion of *implementation* (where a connectionist model implements LOT, say). 'Instantiation' expresses a simple relation between individuals and properties: an individual i instantiates (or is an instance of) a property P if and only if Pi. Where individual computational models or functional

---

For more detailed accounts of the connectionist framework, see, e.g., Rumelhart, McClelland, and the PDP Research Group (1986), McClelland, Rumelhart, and the PDP Research Group (1986), and Smolensky (1988, and MS).

architectures are under consideration, therefore, what is relevant is whether those particular models or architectures *have certain properties* - those, for example, that define the LOT or connectionist frameworks. Thus one might ask of a connectionist architecture whether it has the property of having a combinatorial syntax and semantics, i.e., whether it instantiates, or is an instance of LOT. Or one might ask of an LOT architecture whether its properties also suffice for its being connectionist, i.e., whether that *same LOT architecture* has interconnected processing units that take on various activation values, etc. When one speaks of instantiation, then, a single model or architecture is involved, and what is in question are its properties.

Where *implementation* is at issue, on the other hand, *two* architectures must be considered. A functional architecture FA1 is implemented, if at all, by the execution of a program in a *distinct* functional architecture FA2. Intuitively, the primitive operations, representational structures, etc. of FA1 get "made up" or "constructed" out of the resources of FA2. This is the relation that typically exists, for example, between assembly language architectures, on the one hand, and higher-level architectures like LISP or Pascal, on the other, when the latter are up and running on a computer. With regard to the question whether a particular connectionist architecture *implements* LOT, therefore, what is being asked is *not* whether the connectionist architecture has the property of having a combinatorial syntax and semantics (i.e., whether it instantiates LOT). The issue, rather, is whether it is constructing the resources of some distinct functional architecture with that property.

Notice that in cases of implementation, lower-level architectures typically do not instantiate the characteristic properties of the higher-level ones. An assembly-level architecture implementing LISP, for instance, is not also an instance of LISP: it lacks the necessary primitive resources (e.g., CAR, CDR), and has primitive resources LISP lacks (e.g., various operations on the contents of the accumulator). And, of course, it is also true that an instantiation of LISP need not implement any distinct, higher-level LISP architecture. The notions of instantiation and implementation, therefore, are mutually independent.

There will be a bit more to say of instantiation and implementation in section 3. For now, the point to be emphasized is that it is instantiation, and not implementation, that is involved in the claim of Premise (2) that LOT and connectionism are incompatible. For if LOT is true at the level of cognitive architecture, and connectionist architectures can instantiate LOT, it cannot be inferred that connectionism is false at the level of cognitive architecture. The validity of F&P's argument, therefore, depends on construing the notion of incompatibility in the way I have suggested. (And at any rate, it is agreed on all sides that LOT and connectionism *are* compatible in the sense that connectionist architectures can implement LOT architectures, and vise versa.)

I turn now to (1) and (3).

(1) LOT is true at the level of cognitive architecture

(3) Connectionism is false at the level of cognitive architecture

To say that LOT is true at the level of cognitive architecture is, first, to say this: that the cognitive architecture instantiates LOT; that among all of its properties, one is that it has a combinatorial syntax and semantics. Saying it is true at the *level* of cognitive architecture suggests that there is more than one level of functional architecture, only one of which is the cognitive architecture. This is the familiar picture according to which there are multiple levels of functional architecture, with each architecture being implemented by the architecture one level down, the highest level being the level of cognitive architecture, and the lowest being realized directly in hardware. Premise (1), then, should be understood with this picture in mind, and should be taken as claiming that the level of cognitive architecture instantiates LOT, quite aside from whether functional architectures at any of the other levels do. And (3) says that the properties that define the connectionist framework are *not* instantiated at the level of cognitive architecture, even though, for all we know, they may be at some lower level.[5]

So the issue between LOT and connectionism is over the properties of one particular level of functional architecture - the cognitive architecture. That is the level of psychological explanation, the level it is the aim of cognitive science to

---

[5] The above picture should be taken with a grain of salt, since it is oversimplified, and perhaps also false in its details. Though it is plausible that the highest level of functional architecture is cognitive, for instance, it is not obvious that it must be so. And there is also little reason to suppose there is only one level of cognitive architecture. (This is a common connectionist response to F&P's claim that connectionism will at best implement the cognitive (LOT) architecture, and so can be ignored for psychology. For if the implementation level is cognitive too, psychology cannot ignore it.) For our purposes, however, the situation can be assumed to be more or less as I have described. Doing so will simplify the discussion, and not prejudice matters in any way.

characterize; and the proponents of both frameworks want to claim that level as their own.

On now to F&P's argument.

## 2. Is LOT True at the Level of Cognitive Architecture?

F&P's argument for Premise (1) is meant to be empirical and nondemonstrative. They describe a property of thought they call 'systematicity', and argue that it is best explained on the assumption that the cognitive architecture instantiates LOT. Then, by *inference to the best explanation* (ITTBE), they infer that LOT is true at the level of cognitive architecture.[6]

F&P do not explicitly state that their argument for (1) is an instance of ITTBE, and it is suggested by some things they say that they intend something stronger. I shall discuss these "elements of *a priorism*" in F&P's discussion later in this section. For now, some evidence that ITTBE is indeed what they have in mind is in order.

F&P's argument that LOT is needed to explain the systematicity of cognitive capacities is meant to "precisely parallel" their argument that LOT is required to explain the systematicity of linguistic capacities.[7] However that argument seems

---

[6] F&P also argue for LOT on the basis of their claim that thought is *productive*. But because connectionists are free to deny the productivity of thought, F&P "propose to view the status of productivity arguments for Classical architectures as moot..." (p. 36) I shall follow suit.

[7] F&P (1988, p. 39).

quite clearly to be an instance of ITTBE. For they write,

> There is...a straightforward argument from...the systematicity of language capacity to the conclusion that sentences must have syntactic and semantic structure:...*in effect systematicity follows from the postulation of constituent structure*....
>
> On the view that sentences are atomic, the systematicity of linguistic capacities is a mystery; on the view that they have constituent structure, the systematicity of linguistic capacities is what you would predict. *So we should prefer the latter view to the former.* (p. 38)
>
> (Emphases mine.)

Also, in *Psychosemantics*, where Fodor (1987) offers what appears to be much the same argument for (1), he says,

> I take it that what needs defending here is...*not* the idea that the systematicity of cognitive capacities implies the combinatorial structure of thoughts. I get...[that] claim for free *for want of an alternative account*. (p. 151) (Second emphasis mine.)

I take it that this suffices to show that F&P have *at least* an ITTBE argument in mind. I propose now to examine that argument.

Here is the ITTBE argument for (1):

(.5) Systematicity - a phenomenon to be explained by the cognitive architecture - is best explained by functional architectures that

instantiate LOT

---

(1) LOT is true at the level of cognitive architecture

*Prima facie*, this looks like a perfectly legitimate utilization of ITTBE. I shall argue presently that it is not. First, however, let us consider the phenomenon of systematicity itself, and how LOT is supposed to explain it.

### *2.1 LOT and Systematicity*

F&P introduce the notion of the systematicity of thought thus:

> The easiest way to understand what the systematicity of cognitive capacities amounts to is to focus on the systematicity of language comprehension and production....
>
> What we mean when we say that linguistic capacities are *systematic* is that the ability to produce/understand some sentences is *intrinsically* connected to the ability to produce/understand certain others. You can see the force of this if you compare learning languages the way we really do learn them with learning a language by memorizing an enormous phrase book....[The] point is...that you can learn *any part of a phrase book without learning the rest*....[But you don't], for example, find native speakers who know how to say in English that John loves the girl but don't know how to say in English that the girl loves John. (p. 37)

And a bit further on,

> What does it mean to say that thought is systematic? Well, just as you don't find people who can understand the sentence 'John loves the girl' but not the sentence 'the girl loves John,' so too you don't find people who can *think the thought* that John loves the girl but can't think the thought that the girl loves John. (p. 39)

So the systematicity of linguistic capacities, and cognitive capacities more generally, seems to amount to a particular *relation* among capacities. F&P illustrate this with conditionals of the following form:

> If one has the capacity to say/understand/think that Rab, one also has the capacity to say/understand/think that Rba.

F&P describe this relation among capacities as an "intrinsic connection," but do not say what an intrinsic connection is. Presumably, conditionals of the above form are not meant to be true by logic, for then there would be no explanatory work left for LOT. Nor are they material conditionals since they would be counted as false where the antecedent and consequent are both true by accident, for example. It appears, therefore, that the relation must be understood as a *nomic* one, where the capacities to think that Rab, and to think that Rba, for example, always co-occur as a matter of natural law.[8]

---

[8] This analysis is confirmed by a discussion of Fodor and McLaughlin's (forthcoming) in which they state explicitly that what must be explained are the *lawful* connections among capacities. However, it is still left open by their comments whether systematicity itself requires that the capacities be related

Now how is LOT supposed to explain systematicity? The basic story goes like this. Consider a system that displays systematicity in that Rab can be represented if and only if Rba can (as a matter of psychological law). Suppose the system's functional architecture instantiates LOT in virtue of the following: The representations 'Rab' and 'Rba' are made up of the constituents 'R', 'a', and 'b', and the meanings of 'Rab' and 'Rba' are a function of their structure, and the meanings of the constituents. If we now imagine a computational operation that can construct 'Rab' from 'R', 'a', and 'b' if and only if it can do the same with respect to 'Rba', we are home. But such operations are easy to come by. One whose only constraint on the construction of complex representations is that they have the constituent structure *predicate-name-name* is a case in point. The systematicity of the system's representational capacities, therefore, is explained.

So it looks as though systematicity can be explained by appealing to functional architectures that instantiate LOT. In any event I shall assume that to be so for the purposes of this paper.

### *2.2 The Trouble With F&P's ITTBE Argument*

(.5) Systematicity is best explained by functional architectures that instantiate LOT

---

nomically, or whether mere co-occurrence suffices. This, of course, is a verbal matter, and so is of little consequence. I shall assume in what follows that the nomic component is part and parcel of systematicity.

(1) LOT is true at the level of cognitive architecture

My criticism of this argument is straightforward. We have seen that LOT can account for systematicity. Let us assume it provides the *best* account. It still cannot correctly be inferred by ITTBE that the cognitive architecture instantiates LOT.

Here is why. In the domain of cognitive science, there are many phenomena other than systematicity that must be explained by the cognitive architecture and the algorithms running in it. F&P must consider those phenomena too, and for the following reasons. Suppose there were some non-LOT architecture that explained a particular psychological phenomenon (or perhaps a wide range of phenomena) better than any LOT architecture. Surely, under those circumstances, it would be incorrect to infer that the cognitive architecture instantiates LOT *just because LOT best explains systematicity.* For proponents of the non-LOT architecture would be equally entitled to conclude on the basis of its explanatory success(es) that the cognitive architecture instantiates *it.* Assuming a single cognitive architecture must account for both systematicity and the other phenomenon, however, it cannot be that both LOT *and* the non-LOT architecture are true at the level of cognitive architecture. That is a contradiction. Now because such a scenario is *possible*, it follows that inferring (1) on the basis of (.5) *alone* is a mistake. Further premises are needed.

More generally, here is what has gone wrong. Consider any theory or theoretical framework, T, the aim of which is to explain some domain of

phenomena, D = {P1, P2,...Pn}. And suppose T provides the best explanation of the phenomenon P1. Now ITTBE can be used to infer the truth of T with respect to D *only if P1 is the only element in D*. For otherwise additional premises are required: like (i) that there is no other theory or framework, T*, that is in competition with T to explain the phenomena in D; or (ii) that if there is such a theory or framework, there is no phenomenon, Pi, in D that T* explains better than T; or (iii) that if there is such a phenomenon, T's explanation of P1 is sufficiently better than T*'s explanation of Pi to warrant concluding that T is true and T* is false; and so on. The worry, of course, is that if there is more than one element of D, there may be some other theory T* that does an "overall better job" than T in explaining the phenomena in D, and thus deserves to be counted as true, *even though T offers the best explanation of P1*. In the absence of further premises like (i), (ii), or (iii), therefore, ITTBE can give rise to blatantly false conclusions, as well as to contradictions if applied to incompatible theories.

The upshot is that F&P need something more than (.5) to get (1), since there are phenomena other than systematicity in the domain of cognitive science that must be explained. And because there is a theoretical framework in competition with LOT - connectionism, for example - F&P require a premise along the lines of (ii) or (iii) above. They must argue either that (a) there is *nothing* (non-LOT) connectionist architectures can do better than LOT architectures; or, if there is, that (b) when all the points are tallied, LOT comes out sufficiently ahead of connectionism to warrant accepting LOT as true at the level of cognitive architecture, and connectionism as false.

The prospects for coming up with an argument for either (a) or (b) would appear to be dim. On the surface at least, it seems there *are* things connectionist models do better than LOT models, i.e., phenomena (or aspects thereof) best explained by connectionism, where the connectionist models that do the explanatory work are not instances of LOT. Some examples: error-tolerant content-addressable memory, graceful degradation, flexible and computationally affordable reasoning in terms of prototypes and "emergent schemata", perhaps even *neural realizability*, and so on.[9] If F&P are to argue for (a), therefore, they must argue that none of the above phenomena, or any others, are best explained by connectionist models.

If, on the other hand, F&P agree there are things connectionism does better than LOT, and attempt instead to argue for (b), their prospects for success would seem even dimmer. For unless one of the two frameworks is *obviously* the superior, it is hard to see how F&P could calculate the relative strengths and weaknesses of connectionism and LOT in the absence of a general theory of how to select among competing scientific theories. And of course there is no such theory. Indeed, it would seem the best gauge we have of the relative merits of competing theories in science is in terms of the amounts of support they get by practicing scientists. With regard to LOT and connectionism, however, there is nothing even approaching a general consensus among cognitive scientists. It is too early. The prospects for establishing (b), therefore, seem not very promising at best.

In any event, it is clear that the burden falls squarely upon F&P to provide a

---

[9] For a brief discussion of many of connectionism's attractive features, see Clark (1989, ch. 5). More detailed treatments can be found in Rumelhart and McClelland (1986), McClelland and Rumelhart (1986), and Smolensky (1988).

further argument for either (a) or (b) if they wish to infer that LOT is true at the level of cognitive architecture. That LOT best explains systematicity is simply not enough. I conclude that F&P's ITTBE argument for Premise (1) fails. Of course that is not to say that (1) is false, or even that there could not be a sound ITTBE argument for (1). It is just to say that F&P have provided no reason so far to suppose (1) is true. While they have presented connectionists with the important empirical challenge of explaining systematicity, a challenge connectionists may ultimately be unable to meet, that is no different in kind from empirical challenges facing LOT. Accordingly, LOT gains points, but not victory.

### *2.3 Strains of A Priorism*

At the beginning of this section, I mentioned that certain of F&P's comments suggest that their argument for (1) involves something stronger than ITTBE. Well, it turns out there is an *a priori* argument of sorts in F&P's essay, though not an argument for (1). The conclusion of the argument is that, as a matter of principle, connectionism *cannot* explain systematicity; not just that it *has not* as a matter of empirical fact. (The same argument, essentially, is found in Fodor's *Psychosemantics* (1987), and in a recent paper by Fodor and McLaughlin (forthcoming).[10])

---

[10] In *Psychosemantics*, the argument is directed at "mere intentional realism," of which connectionism is an instance. Here, the argument *does* entail (1), since LOT and mere intentional realism exhaust the range of possible computational frameworks. That is not so, however, when the argument is directed at connectionism, as it is in F&P (1988) and in Fodor and McLaughlin (forthcoming)

Though the argument is typically embedded in the context of a positive (ITTBE) argument for LOT, and often is presented briefly, almost as a side issue, it is striking just how strong the conclusion of the argument is. For, assuming systematicity is a genuine psychological phenomenon, if connectionism *cannot* explain it, *in principle*, connectionism is false at the level of cognitive architecture. Period. Premises (1) and (2), therefore, turn out not to be needed for (3); the *a priori* argument (as I shall call it) gets there directly.

So what *is* the *a priori* argument? Here is a quotation from F&P (1988):

> ...as far as connectionist architecture is concerned, there is nothing to prevent minds that are arbitrarily unsystematic. But that result is *preposterous*....
>
> [Now it's] possible to imagine a connectionist being prepared to admit that while systematicity doesn't *follow from* - and hence is not explained by - connectionist architecture, it is nonetheless *compatible* with that architecture....
>
> [But] it's not enough for a connectionist to agree that all minds are systematic; he must also explain *how nature contrives to produce only systematic minds*. (pp. 49-50)

The embedded argument seems to be this:

(A) For all theories T and phenomena P, a necessary condition for T to be an explanation of P is that T entail P

(B) Systematicity is not entailed by connectionist architecture, since both the presence and absence of systematicity are compatible with it

---

(C) Connectionist architecture cannot explain systematicity

Like F&P's main argument, this too is valid. And it is *a priori* in the sense that the conclusion follows just from claims about the nature of explanation, and the nature of connectionist architecture; excluded is the possibility that a connectionist account of systematicity might be discovered some time in the future. As I shall now argue, however, depending on how one interprets 'connectionist architecture' in (B) and (C), either the argument is sound but uninteresting, or Premise (B) is false. Either way, then, it fails.

'Connectionist architecture', as it appears above, can be taken to refer either to the *connectionist architectural framework*, or to *any possible connectionist model.* If it is the connectionist framework that is meant - the set of properties in virtue of which particular models or functional architectures are or are not connectionist - then the above argument is sound[11] but harmless. For one looks to *particular connectionist models* for explanation, not to the general properties that characterize the connectionist framework. The above argument, understood in this way, therefore, is like the argument that the laws of physics do not explain why a leaf falls as it does, since it is consistent with the those laws that it might have fallen

---

[11] Assuming, as I shall for present purposes, that (A) is true, and that both the presence and absence of systematicity are compatible with the connectionist framework.

some other way. Moreover, the argument applies equally well to the LOT framework: one can easily imagine an LOT model with unconventional computational operations that can construct the representation 'Rab' from its constituents. but not 'Rba'.[12] Now if on the other hand F&P mean *any possible connectionist model* by 'Cognitive architecture', then Premise (B) is false. For, as a rule, each connectionist model will be such that either it displays systematicity or not, and so it will be false that both the presence and absence of systematicity are compatible with it.

The *a priori* argument, therefore, fails to establish F&P's claim that connectionism cannot explain systematicity. It follows that F&P have no argument, demonstrative or otherwise, for their desired conclusion that connectionism is false at the level of cognitive architecture. Both the ITTBE argument for (1), in the context of their main argument, and the *a priori* argument, fail to deliver what F&P want.

## 3. Are LOT and Connectionism Incompatible?

---

[12] F&P (1988) say that in order to construct such an LOT model, one would have to "go out of one's way," though one would not in the case of connectionist models (p. 49). Even if that is true, however, I fail to see what that buys them. And Fodor and McLaughlin (forthcoming) go further, and deny that such LOT models are even possible. They write,

> ...in the Classical architecture, if you meet the conditions for being able to represent aRb, *YOU CANNOT BUT MEET THE CONDITIONS FOR BEING ABLE TO REPRESENT bRa*; the architecture won't let you do so...(p. 17)

That claim simply seems false, so long as primitive operations are included in the specification of the architecture. And they are.

Although F&P's failure to support Premise (1) is sufficient to show that their main argument is unsound, it is worth taking an independent look at their defence of (2), i.e., of the claim that LOT and connectionism are incompatible. For it appears that their argument does not hit the mark there either, and so falters on two counts. I shall first criticize F&P's defense of (2), and then offer some positive reasons for thinking (2) is false. I conclude with some remarks on the implications of the falsity of (2) for the connectionism-LOT debate.

### *3.1 F&P's Defense of (2)*

To claim that LOT and connectionism are incompatible, once again, is to claim that they cannot both be instantiated in the same functional architecture; or what is the same, that there can be no connectionist architecture that is an instance of LOT, and vice versa. Now this claim is a logical or metaphysical one. It is two *sets of properties* - those that define the LOT and connectionist frameworks respectively - that are being declared to be incompatible, just as being a square is incompatible with being a circle. The way one argues for such incompatibility, then, is by showing that nothing could simultaneously satisfy the properties of both sets.

F&P begin in the second section of their paper, by defining LOT in the manner discussed in section 1 of this essay (pp. 3-4): in short, instantiating LOT amounts to having a combinatorial syntax and semantics. Next, they describe a

simple connectionist network that can draw inferences from A&B to either A or B, and contrast that with a description of an LOT machine that can do the same. F&P's argument for the incompatibility of connectionism and LOT, if they have one, is contained somewhere in their discussion of those models. For immediately following that discussion, they embark upon the task of diagnosing how their conclusion might have been missed; in particular, they discuss four ways one might mistakenly think connectionist models instantiate LOT.

I must say I am able to find no explicit argument for (2) in that discussion. Perhaps, however, there is an argument implicit in what they say. If there is, then since they are arguing from a discussion of *particular models*, it would seem that the argument must have more or less the following form:

> From our knowledge of the defining properties of LOT (with the aid of the particular LOT model before us), we see that the connectionist model in the example does not instantiate LOT. That we do, presumably, by noting that the connectionist model has some property, P, that is inconsistent with the defining properties of LOT (viz., a combinatorial syntax and semantics). Finally, we ask whether it is plausible that *all possible* connectionist models have P, i.e., whether P appears to be a necessary condition on connectionist architectures. If and only if it does, (2) can be considered true.[1]

---

[1] As I mentioned above (p. 4), I am hesitant to specify necessary or sufficient conditions on connectionist architectures. Accordingly, it will suffice to establish (2), so far as I am concerned, if F&P offer a *plausible candidate* for a necessary condition on connectionist architectures that is inconsistent with LOT. A property that all *currently existing* connectionist architectures share, for example, will do.

The connectionist network employed in F&P's discussion is illustrated in Figure 1.

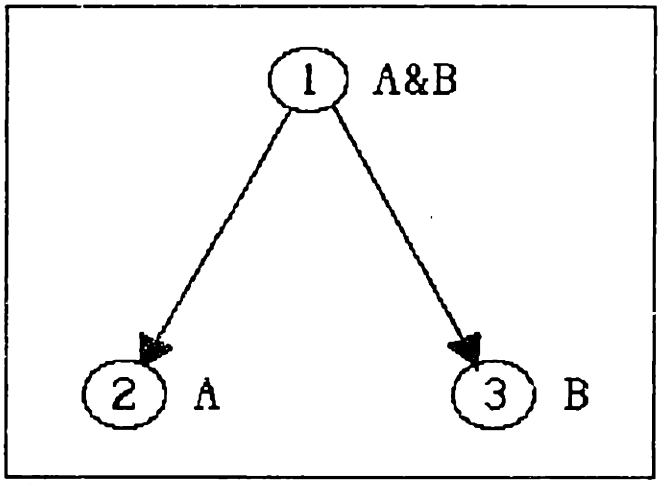


**Figure 1**

Though the diagram looks simple enough, its interpretation is in fact not straightforward. The reason is that instead of the circles representing *units*, as they would in a conventional connectionist network diagram, they represent what F&P call '*nodes*', a node being anything that is either a unit, or an aggregate of units. Consequently, the lines connecting the circles represent either individual weighted connections between units, or multiple weighted connections between aggregates of units. Though I find the notion of a node somewhat puzzling, F&P adopt their terminology, we are told, so their remarks can apply generally both to models that appeal to local representations (realized in single units), and to those that

incorporate distributed representations (realized in aggregates of units).[2]

The "network" portrayed in F&P's diagram contains three nodes, representing the propositions A&B, A, and B, respectively. There are connections from the A&B node to the A node and to the B node. The basic idea underlying the network's operation is that when the A&B node becomes activated, the A node and the B node are both caused to become activated. Thus the network infers both A and B from A&B.

F&P provide a brief description of how a simple Turing machine model that instantiates LOT would make the same inferences, and then they spell out the differences between the connectionist and LOT machines thus:

> In the [LOT]...machine, the objects to which the content A&B is ascribed (viz., tokens of the expression 'A&B') literally contain, as proper parts, objects to which the content A is ascribed (viz., tokens of the expression 'A'). Moreover, the semantics...of the expression 'A&B' is determined in a uniform way by the semantics of its constituents. By contrast, in the connectionist machine none of this is true; the object to which the content A&B is ascribed (viz., node 1) is causally connected to the object to which the content A is ascribed (viz., node 2); but there is no structural (e.g., no part/whole) relation that holds between them. In short, it is characteristic of [LOT]... systems, but not of connectionist systems, to exploit arrays of symbols some of which are atomic...but indefinitely many of which have other symbols as syntactic and semantic parts.... (p. 16)

---

[2] See their footnote 1, p. 5.

It would seem that F&P's argument for (2) must be in that passage, if anywhere; for immediately following it is F&P's diagnosis that I mentioned above. For the sake of clarity, rather than discussing their "network of nodes," I shall discuss, independently, the localist and distributed versions of their network. It will be easier in that way to evaluate whether F&P have made their case for (2).

First the localist version. On that interpretation there is one unit which when activated represents A&B, and which as a consequence causes two other units to become activated, one of which represents A, and the other, B. Now in this case it is patently obvious that the model does not instantiate LOT. And that is because *it has no complex representations*, as required by the definition of 'LOT'. Each of the three representations in the network is realized in an individual unit, and individual units are *structurally simple*. But what follows from that? Well, just that LOT is incompatible with the *localist variety* of connectionism, since every possible localist model is such that each representation in the model is realized by a single, structurally simple unit.

Consider now the distributed version of the network diagrammed above. It is here that things immediately break down. For in the distributed version the details of the model are far too underspecified to determine whether or not it instantiates LOT. In the localist model, the matter is clear: its failure to instantiate LOT is entailed by the fact that each representation is realized by a single unit. So when F&P tell us, as they do in the above passage, that the semantics of the A&B representation is not determined by the semantics of its parts, we know exactly why: it has no parts. But how do we know the same holds when the A&B representation

is understood as realized in a pattern of activity across an aggregate of units? Is it *obvious* that the units of the aggregate could not have semantic values that determine, along with the structure of the aggregate, the semantic value of the whole representation? F&P say that is so, but they do not show it.[3] In the absence of such a demonstration, then, F&P have simply begged the question. It appears, therefore, that they have failed to satisfactorily argue for their claim that connectionism and LOT are incompatible.

### *3.2 Is (2) True?*

Indeed, contrary to what F&P believe, it looks as though distributed models *can* instantiate LOT, *as F&P have defined it.* Not only have they failed to argue for (2), therefore, it looks as though (2) is false. To see this, imagine a distributed representation in some network that is composed of three units which, when activated, have the meanings *John*, *loves*, and *Mary*, respectively. If it is granted that this is a possible scenario, the crucial question is this: Is there a notion of

---

[3] To emphasize the point, the kinds of details needed for the distributed version of the model are like the details we have regarding the localist model. For instance, suppose the distributed model were such that individual units do not have semantic interpretations, and only the patterns of activity across entire aggregates do. If that were a property of the distributed model F&P have in mind, then it would *follow* that the model does not instantiate LOT, since the semantics of the distributed representations are not functions of the semantics of their parts. And the reason would be not that the representations have no parts. They do. Rather, it would be that the parts have no semantic content. Now this property, of course, applies only to a certain class of distributed representations, and so could not suffice to secure (2) for F&P. But it is the *sort* of property that is required, and the sort F&P do not supply.

*structure* applicable to these units, which, when taken with the units' meanings, determines a unique meaning for the distributed representation as a whole - e.g., that John loves Mary rather than, say, that Mary loves John. If there is such a notion, then it seems one has everything that is needed for LOT: There exists a complex representation, the meaning of which is a function of the meanings of its parts and their structure; that is, there is a combinatorial syntax and semantics.

I would submit that the three units along with their meanings *can* have structure in the required sense. For all it is for representational parts to be structured so that a unique meaning, M, is determined is this:

> The parts must have particular (nonsemantic) properties that play a role in causing the machine to behave in ways that *make sense*, given that the complex representation means M.[4]

It seems, however, that can easily be made to hold of the three units in the network we are imagining. The relevant properties of the units would be (some of) their particular connections, and their weights. And the network could be designed so that those properties are causal determinants of certain inferences computed by the network - inferences that make sense (e.g., are valid) on the assumption that the activation of the three units means that John loves Mary.

---

[4] Actually, I intend this as no more than a rough characterization of what it is for representational parts in computational systems to be structured, but as one, also, that is precise enough for my purposes.

Suppose, for example, that as a consequence of the three units being activated, three other units with the meanings *John*, *loves*, and *someone* from some other distributed representation become activated; as do three others from a third distributed representation with the meanings *Mary*, *is-loved-by*, and *someone*. Now given that the meanings of the units from the three distributed representations are fixed,[5] we can make sense of what the network is doing if we assume that our initial distributed representation means that John loves Mary: it is validly inferring that John loves someone, and that Mary is loved by someone. If we assume it means that Mary loves John, however, the network's behavior is unintelligible.

It thus seems quite natural to say that the structural properties among the three units determine, along with the units' meanings, that the distributed representation means *John loves Mary*; and that because the network is causally sensitive to that structure, it is able to make the valid inferences it does. This is just the sort of talk that gets generated with respect to more standard LOT instantiations, like Turing machines that operate on complex representations, for example. However I am at a loss to see any relevant difference between the two kinds of machines in the senses in which they have structured representations.

Consider the sense in which complex representations in Turing Machines are structured, for example. Suppose that upon three squares of a Turing machine's tape there is written, from left to right, 'John', 'loves', and 'Mary'; and suppose the machine is causally sensitive to the fact that the John-square is to the left of the

---

[5] The legitimacy of this assumption is not in contention here. What is in contention is whether the meanings of the individual parts can be *structured*.

Mary-square. Clearly, this does not suffice to determine that the complex representation means John loves Mary, rather than Mary loves John, or vice versa. It is still required further that the *effects of the machine's sensitivity to that fact* be considered in order to determine what the representation means - just as it is with the simple connectionist machine I have described.

Accordingly, I see no reason to deny that that connectionist machine instantiates LOT, as F&P have defined it. It would appear, therefore, that (2) is false.[6] One must be careful, however, not to infer from that that there can be connectionist instantiations of LOT that account for the systematicity of thought. For, as we have seen in section 2, LOT on its own does not suffice to give rise to systematicity; at a minimum, appropriate structure-sensitive operations are required in addition (or, at least they are in the standard sorts of LOT models). Perhaps, also, there are other necessary conditions connectionist architectures will be unable to meet. All this, however, remains to be seen.

### *3.3 Concluding Remarks*

Since it looks as though there can be connectionist architectures that instantiate LOT, we are left with the epistemological possibility that such an architecture might constitute the best available account of the *cognitive architecture*. If that were so, both LOT and connectionism would be true at the level of cognitive

---

[6] For a non-trivial Connectionist network that seems also to instantiate LOT, see Cummins (forthcoming).

architecture. I would like to end this paper by briefly considering what the implications of such a scenario would be for the debate between connectionists and advocates of LOT.

Consider first the question whether there is any sense in which connectionism would have "suffered a defeat," if the cognitive architecture turns out to be a connectionist instantiation of LOT. For after all, an LOT supporter might reason, LOT is true at the level of cognitive architecture, so connectionism cannot really have anything new to offer cognitive science, all claims to the contrary notwithstanding.

Such reasoning, however, is confused. If the cognitive architecture turns out to be an instance of LOT, and the best connectionism can do is implement that architecture, then connectionism *will have* suffered a defeat. (Recall the implementation/instantiation distinction discussed in section 1.) For such a connectionist implementation will be a *mere* implementation of the cognitive architecture in the following two senses: first, it will be irrelevant for the purposes of psychological explanation, in that it can be ignored[7]; and second, although it might be true of how the cognitive architecture is implemented *in fact*, it will still be possible that the cognitive architecture can be implemented in many different ways, just as LISP or Pascal can.

But the issue at hand has nothing to do with whether connectionist models can, at best, *implement* the cognitive architecture (which for the time being we are assuming is an instance of LOT). The issue, rather, has to do with whether the

---

[7] But see fn. 5.

cognitive architecture *is* a connectionist architecture, one which happens also to instantiate LOT. If it is, then connectionism is *crucially* relevant to psychological explanation. For the explanations of psychological phenomena will be connectionist explanations through and through; not explanations in terms of some other LOT architecture like production systems, for example. Connectionism, therefore, will hardly have been defeated.

There is one sense, however, in which connectionism, or more precisely the views of certain connectionists, will have suffered a defeat under the circumstances we are imagining. I have in mind the many claims in the connectionist literature about psychological phenomena that can be explained by connectionist models without appeal to symbol systems, complex representations, LOT, etc. If LOT is needed for those explanations, such claims will of course be false. This kind of defeat, however, is not serious. For, as we have seen, it does not imply in the least that connectionism can be ignored for the purposes of psychology, that it has nothing new and interesting to offer, that it ought to be denied funding, or anything else of the sort.

It should be clear by now that even if connectionism attempts to model the cognitive architecture by instantiating LOT, that in no way excludes the possibility that connectionist architectures might differ in interesting and important ways from all other LOT architectures developed thus far. It should also be clear just how abstract an architectural framework LOT is. Even if the cognitive architecture turns out to instantiate LOT, the vast majority of psychological phenomena will properly be explained by far more specific details than those embodied in the definition of LOT.

## REFERENCES


Clark, A. (1989). *Microcognition.* Cambridge: MIT Press.

Cummins, R. (forthcoming). "The Role of Representation in Connectionist Explanations of Cognitive Capacities," to appear in D. Rumelhart, W. Ramsey, and S. Stich (eds.) *Philosophy and Connectionist Theory.*

Fodor, J.A. (1987). *Psychosemantics.* Cambridge: MIT Press.

Fodor, J.A., & McLaughlin, B. (forthcoming). "Connectionism and the Problem of Systematicity: Why Smolensky's Solution Doesn't Work," to appear in *Cognition.*

Fodor, J.A., & Pylyshyn, Z.W. (1988). "Connectionism and Cognitive Architecture: A Critical Analysis," *Cognition* **28**, 3-71.

Haugeland, J. (1985). *Artificial Intelligence.* Cambridge: MIT Press.

McClelland, J.L., Rumelhart, D.E., & the PDP Research Group (eds.). (1986). *Parallel Distributed Processing*, Vol 2. Cambridge: MIT Press.

Pylyshyn, Z.W. (1984). *Computation and Cognition.* Cambridge: MIT Press.

Rumelhart, D.E., McClelland, J.L., & the PDP Research Group (eds.). (1986). *Parallel Distributed Processing*, Vol 1. Cambridge: MIT Press.

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

Smolensky, P. (MS). "Representation in Connectionist Networks."