CANADIAN JOURNAL OF PHILOSOPHY
Volume 26, Number 1, March 1996, pp. 101-122

Searle, Syntax, and Observer Relativity

RONALD P. ENDICOTT
Arkansas State University
State University, AR 72467-1890
USA

In his book *The Rediscovery of the Mind* (hereafter *RM*), John Searle attacks computational psychology with a number of new and boldly provocative claims.[1] Specifically, in the penultimate chapter entitled 'The Critique of Cognitive Reason,' Searle targets what he calls 'cognitivism,' according to which our brains are digital computers that process a mental syntax. And Searle denies this view on grounds that *the attribution of syntax is observer relative*. A syntactic property is arbitrarily assigned to a physical system, he thinks, with the result that syntactic states 'do not even exist except in the eyes of the beholder' (*RM*, 215). This unabashed anti-realism differs significantly from Searle's earlier work. The Chinese room argument, for example, was intended to show that syntactic properties will not suffice for semantics, where the syntax was realistically construed.[2] But now Searle claims that physical properties will not suffice to determine a system's syntactic properties. He puts the point in terms of what is 'intrinsic to physics':

[1] John Searle, *The Rediscovery of the Mind* (Cambridge, MA: The MIT Press 1992). The material in question (chap. 9) is reprinted almost verbatim from Searle's Presidential Address 'Is the Brain a Digital Computer?' *Proceedings and Addresses of the American Philosophical Association* **64** (1990) 21-37. Note that the arguments are directed against 'classical' cognitivism. Searle's stand on connectionism appears to have changed. Cf. his 'Is the Mind's Brain a Computer Program?' *Scientific American* **262** (1990) 26-31, with *The Rediscovery*, 246-7.

[2] John Searle, 'Minds, Brains and Programs,' *The Behavioral and Brain Sciences* **3** (1980) 417-24

> This is a different argument from the Chinese room argument, and I should have seen it ten years ago, but I did not. The Chinese room argument showed that semantics is not intrinsic to syntax. I am now making the separate and different point that syntax is not intrinsic to physics. For the purposes of the original argument, I was simply assuming that the syntactical characterization of the computer was unproblematic. But that is a mistake (*RM*, 210).[3]

So Searle had assumed that the assignment of syntax was relatively unproblematic. But no more. Syntactic states are infected with observer relativity so that, ontologically speaking, they are mere shadows cast by outside observers. This is a provocative claim indeed. Of course, others have pressed the issue of observer relativity before, but at the level of semantic interpretation.[4] Searle's arguments, on the other hand, are leveled at the syntax, and consequently draw upon different considerations that are worthwhile to set forth and examine in detail. Nevertheless, I will show that these new claims cannot withstand careful scrutiny, and that when this material is sifted through and its premises laid bare, the arguments are found to rest upon unsupported and erroneous assumptions.

I Arbitrariness & Universal Realizability

I will focus on two basic lines of support for observer relativity that Searle provides. The first will be examined in this section, though it is not in my view the most serious of the two. The second, to be discussed later, raises some issues that deserve careful attention, more so, I suspect, than Searle has given them. But first the lesser point. *Syntax is observer relative because the assignment of syntax is arbitrary, leading to the universal realizability of computers.* That is, Searle begins his attack on cognitivism by making

[3] Let me warn that Searle has an exasperating proliferation of meanings for the term 'intrinsic.' Here it means (a) *determination* by a set of lower-level properties, i.e., syntax is not intrinsic to physics because physics will not suffice to determine syntactic properties. But elsewhere it means (b) *definability* in terms of a particular class of predicates, i.e., syntax is not intrinsic to physics because it is 'not defined in terms of physical features' (*RM*, 225). But more often than not, Searle uses 'intrinsic' to mean (c) *real*, ontologically speaking, and thus mark the distinction between 'the real thing' as opposed to the merely 'as if,' 'derived,' or 'observer-relative' (78-80, 208).

[4] W.V.O. Quine, *Word and Object* (Cambridge, MA: The MIT Press 1960); also Quine's 'On the Reasons for Indeterminacy of Translation,' *Journal of Philosophy* **67** (1970) 178-83; and Daniel Dennett, 'Reflections: Real Patterns, Deeper Facts, and Empty Questions,' in *The Intentional Stance* (Cambridge, MA: The MIT Press 1987) 37-42.

heavy weather out of the fact that an 'assignment' or 'interpretation' must be made. After alluding to the notion of a Turing machine, for example, he expresses some puzzlement over the fact that 'To find out if an object is really a digital computer, it turns out that we do not actually have to look for 0's and 1's, etc.; rather we just have to look for something that we could *treat as* or *count as* or *could be used to* function as 0's and 1's' (*RM*, 206). And, again, Searle claims that the crucial difference between functional properties like being a carburetor or a thermostat and the suspect syntactic properties is that the latter involve an assignment to a physical system: 'The classes of carburetors and thermostats are defined in terms of the production of certain *physical* effects. That is why, for example, nobody says you can make carburetors out of pigeons. But the class of computers is defined syntactically in terms of the *assignment* of 0's and 1's' (*RM*, 207).

Yet it is clear that the emphasis should not be upon the mere fact that we make an assignment to or treat something as a 0 or 1. Consider the progression from a little girl treating her doll as if it were a person; the young woman treating her cat as if he were a person; and the graduate student treating her adviser as if *he* were a person. Some treatments may be correct, and this depends entirely upon the attitude and objects concerned. Hence Searle's remarks should be understood as an attempt to underscore the *arbitrary nature* of the attribution of syntax, so much so that he thinks virtually *anything* will count as a computer. This is the point about universal realizability. He says:

> The same principle that implies multiple realizability would seem to imply universal realizability. If computation is defined in terms of the assignment of syntax then everything would be a digital computer, because any object whatever could have syntactical ascriptions made to it. You could describe anything in terms of 0's and 1's ... we wanted to know how the brain works, specifically how it produces mental phenomena. And it would not answer that question to be told that the brain is a digital computer in the sense in which stomach, liver, heart, solar system, and the state of Kansas are all digital computers. (*RM*, 207-8)

Certainly if the cognitivist's understanding of computation requires that we view everything as a computer, even the state of Kansas, then it must be rejected out of hand. But does it?

II Constraints on Interpretation

The assignment of syntax need not be arbitrary, or lead to any universal realizability, if there are additional constraints at work within our interpretive practice beyond the mere assignment of syntax that will isolate a limited set of physical systems and count only those systems as genuine computational devices. To this end, the cognitivist will appeal to familiar considerations about the functional architecture of the system in question, its causal and historical connections with the environment, as well as the intentional behavior we are trying to explain by the postulation of syntactic structures which can then be semantically interpreted in such a way as to capture important psychological generalizations. Consider the state of Kansas. It does not behave *at all* (especially if it is an abstract object), and so there can be no question about postulating syntactic states with semantic content in order to explain its behavior. In contrast, it is often maintained that our own linguistic behavior is systematic and productive in a way that *requires* a language of thought.[5] I do not say these arguments are absolutely decisive. But these are the kind of points upon which any further debate must turn.

Or consider the appeal to functional architecture. Let the physical states of a system be those designated as causally relevant by the appropriate physical theory (i.e., no metaphysical gerrymandering); and let the formal states be those designated as functionally relevant by the appropriate computer program. We can then say that a system is a genuine computational device when there is a correspondence between its physical states and its formal states such that the causal structure of the physical system is isomorphic to the formal structure of the computational operations. Consequently, for any given computational operation there will be an indefinite number of physical systems, like the state of Kansas, which simply *fail to have the right causal structure*.[6] Clearly Searle takes these constraints too lightly. He says:

> For any program there is some sufficiently complex object such that there is some description of the object under which it is implementing the program. Thus for example the wall behind my back is right now implementing the Wordstar program,

[5] Jerry Fodor, *The Language of Thought* (Cambridge, MA: Harvard University Press 1975); Fodor's appendix to *Psychosemantics* (Cambridge, MA: The MIT Press 1987); and a more cautious discussion in Andy Clark, *Microcognition* (Cambridge, MA: The MIT Press 1989), ch. 8.

[6] An anonymous referee has suggested that we might strengthen the point against Searle by distinguishing between 'implementational capacity' and 'implementation.' The former is characterized by the mathematical notion of isomorphism whereby the physical states of the system are isomorphic to the transitions between states that the program specifies. And here it may be conceded that a given system has the 'implementational capacity' with respect to a large number of programs. Yet this is not to say that the system actually 'implements' all those programs. For actual implementation is determined by the appropriate causal interactions between the physical store of the program and the physical device itself.

> because there is some pattern of molecule movements which is isomorphic with the formal structure of Wordstar. But if the wall is implementing Wordstar then if it is a big enough wall it is implementing any program, including any program implemented in the brain. (*RM,* 208-9)

But unless we allow for miracles, the wall is not implementing Wordstar if we consider *non-gerrymandered* physical units, or if we consider not an isolated time slice but a *physical system whose parts have the disposition to causally interact in the way specified by the program.* (If walls were like Searle says they are, we could chop them up in little pieces, sell them as the latest microprocessors, and become extremely wealthy! But the consumer won't buy even if we let Searle make the pitch.) Thus, barring well-worn controversies which surround recherché cases like Block's Chinese Nation,[7] the cognitivist could say there is a reasonably clear set of systems which, given their distinctive features, make it compelling to believe that they actually instantiate the properties which computational psychology attributes to them in a fully realistic, nonobserver-relative way.[8]

Curiously enough, Searle mentions the role of such additional constraints, and even seems to concede the point at issue, but then repeats his charge about observer relativity:

> I think it is possible to block the result of universal realizability by tightening up our definition of computation. Certainly we ought to respect the fact that programmers and engineers regard it as a quirk of Turing's original definitions and not as a real feature of computation. Unpublished works by Brian Smith, Vinod Goel, and John

[7] Ned Block, 'Troubles with Functionalism,' in Ned Block, ed., *Readings in Philosophy of Psychology* 1 (Cambridge, MA: Harvard University Press 1980) 268-305; but cf. P.S. Churchland and P.M. Churchland, 'Functionalism, Qualia, and Intentionality,' in J.I. Biro and Robert W. Shahan, eds., *Mind, Brain, and Function* (Norman: University of Oklahoma Press 1981) 121-45; and William Lycan, 'Form, Function, and Feel,' *Journal of Philosophy* **78** (1981) 24-50.

[8] Or consider again the point about behavior. The remarkable fact about systems like ourselves is that we behave *rationally,* maximizing truth. A further constraint is therefore that any syntactic attribution lead to rationality and truth. So when we discover an interpretation of some 'perceived' syntactic objects which makes sense of the system and its behavior, and if those syntactic patterns continue in such a way that preserves truth, rationality, and the overall fit of the system with its environment, then our hypothesis is empirically well grounded. Notice that this coalesces nicely with my point in the text that we should consider, not a time slice of the wall considered in isolation, but a physical system whose structure has parts with the disposition to causally interact. For rationality concerns the manifestation of such dispositions over time, the coordination of beliefs with action, in short, a continuing system's adjustment of intentionality to the world.

> Batali all suggest that a more realistic definition of computation will emphasize such features as the causal relations among program states, programmability and controllability of the mechanism, and situatedness in the real world. All these will produce the result that the pattern is not enough. There must be a causal structure sufficient to warrant counterfactuals. But these further restrictions on the definition of computation are no help in the present discussion *because the really deep problem is that syntax is essentially an observer-relative notion.* (*RM*, 209)

Fair enough. Causal relations, programmability and control, and situatedness of the system will block the result of universal realizability. Then *why* does Searle repeat his charge? He continues:

> *The multiple realizability of computationally equivalent processes in different physical media is not just a sign that the processes are abstract, but that they are not intrinsic to the system at all. They depended on an interpretation from the outside.* We were looking for some facts of the matter which would make the brain processes computational; but given the way we have defined computation, there never could be any such facts of the matter. We can't, on the one hand, say that anything is a digital computer if we can assign a syntax to it, and then suppose there is a factual question intrinsic to its physical operation whether or not a natural system such as the brain is a digital computer. (*RM*, 209-10)

But something is amiss, an error which shows itself in the last sentence. For the cognitivist does *not* say 'anything is a digital computer if we can assign a syntax to it,' pure and simple, and then go on to 'suppose there is a factual question intrinsic to its physical operation' about whether the brain is a computer. Rather, the cognitivist says that anything is a computer *if we can assign a syntax to it under the conditions specified by the additional constraints on interpretation*, those to which Searle had already made concession, and in light of which there *is* a fact of the matter which distinguishes the class of things that are intrinsically computers from Searle's much wider set of universally realized and observer-relative computers. Compare in this regard a closely parallel debate over Quine's indeterminacy of translation.[9] It may be correct to say the meaning of 'gavagai' is indeterminate with respect to condition C (behavioral evidence). But it certainly does not follow, though it may be true, that 'gavagai' is indeterminate with respect to *other conditions* C* (behavioral evidence plus intentions of the speaker plus neurophysical similarity,

[9] See Donald Davidson and Jaakko Hintikka, eds., *Words and Objections* (New York: Humanities Press 1969). Quine and others would reject the implication that indeterminacy in radical translation arises solely from behaviorist constraints. This is a large question I cannot address here. But however it is to be resolved, I assume Searle is not simply extending Quine's argument to the area of syntax. Cf. Searle, 'Indeterminacy, Empiricism, and the First Person,' *Journal of Philosophy* **84** (1987) 123-46.

etc.). In the same way, syntax may be observer relative with respect to an assignment of zeros and ones to physical variables *tout court*, but not with respect to an assignment under the constraints of rich behavioral repertoire plus non-gerrymandered causal structure and the like.

Finally, it should be clear that, without further argument, Searle simply cannot press the above quoted claim that 'the multiple realizability of computationally equivalent processes in different physical media is not just a sign that the processes are abstract, but that they are not intrinsic to the system at all' (*RM*, 209). For multiple realization is not a 'sign' of any such thing, seeing that *there exist multiply realized types that are not observer relative*. Consider those multiply realized types within the domain of physical or natural science, for example.[10] Hence multiple realizability does not translate into universal realizability, nor does it indicate anything about the status of the property which happens to enjoy such plasticity, whether it be intrinsic or relational, real or observer relative, or whatever. Searle's remarks therefore appear question-begging, a mere reassertion of an otherwise and to this point unsupported claim about observer relativity. So let us turn to the more interesting argument.

III The Homunculus Fallacy

Searle's second major line of argument is that the *cognitivist's view of syntax is observer relative because it involves the homunculus fallacy*. He begins in this way:

> Most of the works I have seen in the computational theory of mind commit some variation on the homunculus fallacy. The idea is always to treat the brain as if there were some agent inside it using it to compute with. A typical case is David Marr ... who describes the task of vision as proceeding from a two-dimensional visual array on the retina to a three-dimensional description of the external world as output of the visual system. The difficulty is: Who is reading the description? (*RM*, 212)

Notice that the difficulty, just hinted at here, is not merely that the cognitivist postulates an inner interpreter to explain mental processes — the complaint that the intentional behavior of a system (perceiving the world, understanding English, etc.) is explained *in terms of other intentional behavior* (the intentional behavior of 'sub-personal' systems or

[10] See my 'On Physical Multiple Realization,' *Pacific Philosophical Quarterly* **70** (1989) 212-24.

agencies like input analyzers, sentence parsers, etc.), leaving us in the very place we began, theoretically speaking. No, Searle thinks there is something objectionable *in the way* the homunculus is postulated, at a level where there can be no interpreter, no one to 'read the description.' This point surfaces a bit more clearly when Searle considers the view that all homunculi are 'discharged' by explaining higher-level computations in terms of more basic functions.[11] He describes the strategy in this way:

> Many writers feel that the homunculus fallacy is not really a problem because, with Dennett ... they feel that the homunculus can be "discharged." The idea is this: Because the computational operations of the computer can be analyzed into progressively simpler units, until eventually we reach simple flip-flop, "yes-no," "1-0" patterns, it seems that the higher-level homunculi can be discharged with progressively stupider homunculi, until finally we reach a bottom level of simple flip-flop that involves no real homunculus at all. The idea, in short, is that recursive decomposition will eliminate the homunculus. (*RM*, 212-13)

What, then, is the problem? Searle continues:

> At the bottom level are a whole bunch of homunculi who just say "Zero one, zero one." All of the higher levels reduce to this bottom level. Only the bottom level really exists; the top levels are all just *as if*. Various authors ... describe this feature when they say that the system is a syntactical engine driving a semantical engine. But we still must face the question we had before: What facts intrinsic to the system make it syntactical? What facts about the bottom level or any other level make these operations into 0's and 1's? *Without a homunculus that stands outside the recursive decomposition, we do not even have a syntax to operate with.* The attempt to eliminate the homunculus fallacy through recursive decomposition fails, because the only way to get the syntax intrinsic to the physics is to put a homunculus in the physics. (*RM*, 213-14)

Unfortunately, Searle is not terribly explicit about how this argument proceeds, and there is much in this passage that stands in need of clarification, even correction.[12] However, Searle seems to be gesturing

[11] See Daniel Dennett, 'Why the Law of Effect Will Not Go Away,' rpt. in *Brainstorms* (Cambridge, MA: The MIT Press 1978) 71-89; and Marvin Minsky, *The Society of Mind* (New York: Simon & Schuster 1985).

[12] Searle is wrong to say that the cognitivist believes 'only the bottom level really exists; the top levels are all just *as if*.' For the standard picture is that the upper levels exist with the same robust sense of reality, only they supervene on the lower levels. See John Haugeland, 'Semantic Engines: An Introduction to Mind Design,' in Haugeland, ed., *Mind Design* (Cambridge, MA: The MIT Press 1981), 1-34. Indeed, Haugeland states a view *in exact opposition* to that which Searle attributes to the cognitivist (more curious still, since Searle refers to Haugeland's piece! [*RM*, 213]). Haugeland says: '*Once we see this point, we see how foolish it is to say that computers are*

towards a tension involving *the connection between symbols and interpreters versus the elimination of such interpreters via a recursive decomposition of computational tasks*. It is the connection between symbol and interpreter that moves Searle to ask his rhetorical question a few quotations back: 'Who is reading the description?' and it is that same connection which moves him in the above passage to claim that 'without a homunculus ... we do not even have a syntax to operate with.' Moreover — and this is the crux of the argument — it is in light of this supposed connection that Searle finds the cognitivist's recursive decomposition problematic. Intuitively speaking, *the connection between symbol and interpreter is severed when the cognitivist 'eliminates' the homunculus*. Hence, though Searle offers no formalization, perhaps the best way to construe his reasoning is as follows:

(i) If there are symbols, there must be an interpreter for those symbols.

(ii) There are no interpreters for the primitive symbols within computational systems.

(iii) So there are no primitive symbols within computational systems.

And what holds for the primitive symbols, the zeros and ones of classical architectures, holds *mutatis mutandis* for any other symbols generated by them. Syntax does not exist, save in the eye of the beholder.

Clearly Searle is committed to (i), the connection between symbols and interpreters, as shown already. Indeed, speaking of the language of thought, he says plainly: 'something is a sentence only relative to some agent or user who uses it as a sentence' (*RM*, 210). And Searle affirms (ii), that given the cognitivist's overall position, there can be no interpreter for the basic symbols of computational theory. We cannot 'put a homunculus in the physics' he says; yet all higher-level homunculi have been explained away. For, again, as Searle views the cognitivist's strategy, the 'recursive decomposition will *eliminate* the homunculus' (*RM*, 213, my italics). But why accept this argument?

nothing but great big number crunchers, or that all they do is shuffle millions of "ones" and "zeros." Some machines are basically numerical calculators or "bit" manipulators, but most of the interesting ones are nothing like that.... The machine one cares about — perhaps several levels of imitation up from the hardware — may have nothing at all to do with bits or numbers; and that is the only level that matters' (14-15).

IV Symbols, Interpreters, and Homunculi

Consider premise (ii), which says that there are no interpreters for the primitive symbols within computational systems. Why not say there *are* interpreters of the requisite sort, namely, ourselves? In other words, grant, purely for the sake of argument, that the 'sub-personal' level homunculi have been eliminated. Still, we at the personal level remain, and we see that certain physical patterns are correctly interpreted as primitive symbols, as a syntax of zeros and ones, upon which the relevant computations will occur. Furthermore, Searle must agree about this 'homunculus that stands outside the recursive decomposition,' since any claim about observer relativity *entails* such an observer. Of course, Searle believes that we stand to the symbols of computational theory in the alleged observer-relative way. But that is precisely the point to be proved, not assumed in support of the argument's key premise.

Also, premise (ii) betrays a deep misunderstanding of the cognitivist's position. Recursive decomposition into simpler computational tasks does not 'eliminate' the homunculi, as Searle maintains. Quite to the contrary, the strategy either *reductively explains* them, or, what is more fitting for the cognitivist, it *nonreductively explains* all computational tasks in terms of lower-level physical processes by a doctrine of supervenience.[13] Either way, ontology is *preserved*, not eliminated. The phenomena described in the upper reaches of computational theory exist with the same robust sense of reality as water vis-à-vis H_2O (preservation through reduction), or being a planet vis-à-vis the appropriately diverse material compositions (preservation through nonreductive supervenience). So it is only by conflating the eliminitivist approach with its realist competitors that Searle can affirm (ii).

Does premise (i) fare any better? Not at all. Premise (i) says that if there are symbols, there must be an interpreter for those symbols.[14] But why

[13] That is, cognitivists are typically anti-reductionists, and supervenience is the anti-reductionist's doctrine of choice. See Jaegwon Kim, 'Concepts of Supervenience,' *Philosophy and Phenomenological Research* **45** (1984) 153-76; and Terence Horgan, 'From Supervenience to Superdupervenience: Meeting the Demands of a Materialist World,' *Mind* **102** (1993) 555-86. This is not to say, however, that all cognitivists would accept a claim about supervenience. Dennett, e.g., maintains that computational patterns exist in some sense but are indeterminate inasmuch as they depend upon adopting the intentional stance ('Reflections: Real Patterns, Deeper Facts, and Empty Questions,' 39-40).

[14] Parenthetically, (i) must be understood aright. It does *not* mean: 'If there are mental symbols, there must be a *mind* for those symbols,' which might be true simply on grounds that a mental state cannot exist apart from a mind in which the state inheres.

believe that? After all, the role of an interpreter is conceptually linked to public languages whose symbols are conventionally associated with their intentional objects, all to aid in the communication between individuals. It is only insofar as we mean or intend something by using a symbol in a particular way that the role of an interpreter cannot be eliminated. Yet matters are quite different for symbols in any alleged language of thought since, by hypothesis, the cognitivist is proposing a theory of 'natural' or 'nonderivative' meaning whereby we do not literally mean anything at all. The syntax is not *used by us* but *occurs in us* and has the meanings it does along the lines suggested by some appropriate theory of mental content, whether it be causal covarience, functional role, adaptational role, or perhaps something else altogether.[15]

Moreover, following Dennett, we may buttress the cognitivist's position by comparing the biologist's talk of DNA codes.[16] Very simple biological structures can read codified information without the aid of any homunculus, a little man inside the sperm cell, as was once postulated. So there can be representational states *in some sense* without anyone to interpret them, and the cognitivist's syntactic structures could be viewed in the same way. And surely the cognitivist's symbolic codes *can* be understood in a similar fashion, independent of any full-blooded interpreters as intended by premise (i). For remember that the homunculi are not really agent-like entities on the model of personal level language users. The cash value of the metaphor is a realism, yes, but about *intentionally specified states and processes*. And it is such states and processes that are broken down into progressively simpler units until we reach the level of implementation, the moving about zeros and ones, or the opening and closing of logic gates, which can then be explained in purely mechanistic terms.

Put in a different way, when we move from personal substance terms like 'perceiver' and 'interpreter' to process terms like 'perceptual encoding at the periphery of the input system,' and when move from high-level to low-level descriptions of those computational processes, then the temptation to saddle the cognitivist with a primitive-level agent-like interpreter should all but disappear. Compare in this regard Searle's

No, there is an additional function which (i) is intended to capture, i.e., that the mind must not only 'have' but 'interpret' its (symbolic) states.

[15] See Robert Cummins, *Meaning and Mental Representation* (Cambridge, MA: The MIT Press 1989).

[16] Dennett, 'Evolution, Error, and Intentionality,' in *The Intentional Stance*, 102-3

previously quoted remark about David Marr, whose theory of vision describes an algorithm which transforms two-dimensional visual arrays into three-dimensional descriptions. Searle's rhetorical question was then: '*Who is reading the description*?' (*RM*, 212, my italics) But if we choose a description, not from the upper-reaches of Marr's level of 'algorithm,' but one gleaned from his more basic 'implementation' level, and if we are aware of the literal states and processes implied by the computational model, then Searle's type of question will appear inane, like asking: '*Who opened the logic gate*?'

These are just a few reasons why the cognitivist should in nowise accept premise (i). Does Searle, then, provide any positive, nonquestion-begging support for the connection between symbols and interpreters? Ultimately, no. We find nothing further in his penultimate chapter, 'The Critique of Cognitive Reason,' which contains the arguments canvassed thus far.[17] Therefore, as it stands, Searle's advertised critique of cognitivism must be judged a failure. All the same, perhaps the critique need not stand alone. There is one last place to look.

V A Possible Consciousness Connection

In an earlier chapter entitled 'The Unconscious and Its Relation to Consciousness,' Searle defends a closely related connection principle according to which genuine intrinsic 'unconscious intentional states are in principle accessible to consciousness' (*RM*, 156). In fact, Searle takes this to be a mark of the mental quite generally, that 'mental states are candidates for consciousness' (*RM*, 161), a basic theme which dominates Searle's *Rediscovery*. Consequently, on the suggestion before us, *the aforementioned connection between symbols and interpreters is but an instance of this more general connection between intrinsic mental states and the con-*

[17] Admittedly, ch. 9 does contain two additional claims: that 'syntax has no causal powers' (*RM*, 214-22); and that 'the brain does not do information processing' (*RM*, 222-5). Nevertheless, those claims hinge upon what has gone before. The reason *why* syntax has no causal powers is that syntax is observer relative: 'But the difficulty is that the 0's and 1's as such have no causal powers *because they do not exist except in the eyes of the beholder*' (215, my italics). And the reason *why* the brain does not process information is that information processing is observer relative: 'The computer then goes through a series of electrical stages that the *outside agent* can interpret both syntactically and semantically even though, of course, *the hardware has no intrinsic syntax or semantics: It is all in the eye of the beholder*' (223, my italics).

sciousness of an interpreter.[18] So we would do well to finish by looking at Searle's material on consciousness.

Yet there is trouble from the start. For it would already seem that the points raised in the previous section will tell against any substantive consciousness connection principle. If mental symbols need not have interpreters, they need not have any interpretive acts of consciousness, someone to be consciously aware of them.[19] Consider again DNA codes, or the logic gates at the basic level of cognition. Hence, in light of the foregoing, we should expect some fairly persuasive support for Searle's principle, something that would, if sound, outweigh the criticism gathered thus far. Unfortunately we are apt to be disappointed. Searle claims to have an 'argument,' though not 'a simple deduction from axioms' (*RM*, 155-6). It appears as an intriguing list of claims, the most important of which are these:

(1) *The Irreducibility of Intentionality*: 'The aspectual feature [i.e., intentionality, how only some "aspects" of the intentional object are presented to thought] cannot be exhaustively or completely char-

[18] My thanks to an anonymous referee for suggesting a connection between the two connection principles. It should be underscored that this is *only* a suggestion, one that Searle nowhere explicitly makes. But we do find passing reference to consciousness in his later critique from ch. 9 (see *RM*, 219-20). Also, we could, if we wish, piece together an attack on cognitivism with the aid of this connection principle that parallels argument (i) through (iii), viz.: *(i*) If there are intrinsic mental states, they are in principle accessible to consciousness. (ii*) The syntactic states of computational theory are not in principle accessible to consciousness. (iii*) So the syntactic states of computational theory are not intrinsic mental states.* And, although Freudian theory is the *designated* target of Searle's attack in ch. 7 on 'deep' unconscious states (151, 167-73), he does cite Jackendorff's distinction between the 'computational mind' and the 'phenomenological mind' as an 'extreme version' of an approach which emphasizes the importance of such unconscious facts (152). Cf. his earlier version of the same material on consciousness specifically tailored to cognitivism, in 'Consciousness, Explanatory Inversion, and Cognitive Science,' *Behavioral and Brain Sciences* **13** (1990), 589.

[19] This is true if consciousness is understood in terms of 'awareness,' which Searle admits is a near synonym (*RM*, 84), since awareness is arguably an interpretive act. If, on the other hand, we understand consciousness in the phenomenal 'what it is like' sense, then, as several authors have pointed out, nothing follows about the status of the cognitivist's mental syntax. In particular, we cannot rule out the possibility that syntactic states have a certain phenomenal feel associated with them. See Ned Block, 'Consciousness and Accessibility,' *Behavioral and Brain Sciences* **13** (1990), 597; and cf. a similar point by Daniel Dennett in his review of Searle's book, *Journal of Philosophy* **90** (1993), at 197-8, where Dennett speculates that the phenomenal feel may be completely cut off from any process of the subject's first-person awareness.

acterized solely in terms of third-person, behavioral, or even neurophysiological predicates' (*RM*, 157-8);

(2) *The Exclusive Neurophysical Ontology for the Unconscious*: 'But the ontology of unconscious mental states, at the time they are unconscious, consists entirely in the existence of purely neurophysiological phenomena' (*RM*, 159);

(3) *Unconscious Thoughts as Possibilia*: 'The notion of an unconscious intentional state is the notion of a state that is a possible conscious thought or experience' (*RM*, 159);

(4) *The Causal-Dispositional Analysis of Unconsciousness*: hence 'the ontology of the unconscious consists in objective features of the brain capable of causing subjective conscious thoughts' (*RM*, 160).[20]

Now I assume (3) and (4) are inferences to the best explanation. That is to say, given (1) that intentionality or what Searle calls 'aspectual shape' is not reducible to other terms, including neurophysical terms, and given (2) that the only facts which answer to the unconscious states of an individual are completely neurophysical in nature, then it is plausible to believe (3) that unconscious intentional states are but the mere possibilities of this neurophysiology, being (4) the disposition of the same neurophysiology to cause conscious states that reflect the intentional features ascribed at the unconscious level. At least this seems to be Searle's logic. He summarizes the argument in this way:

> When you make a claim about unconscious intentionality, there are no facts that bear on the case except neurophysiological facts. There is nothing else there except neurophysiological states and processes describable in neurophysiological terms. But intentional states, conscious or unconscious, have aspectual shapes, and there is no aspectual shape at the level of neurons. So the only fact about the neurophysiological structures that corresponds to the ascription of intrinsic aspectual shape is the fact that the system has the causal capacity to produce conscious states and processes where those specific aspectual shapes are manifest. (*RM*, 161)

[20] Propositions (1) through (4) correspond to Searle's 4. through 7. respectively. Searle's first three, which I omit in the text, simply mark the distinction between 'intrinsic' versus 'as if' intentionality, locate the class of unconscious mental states within the intrinsic, and then define the essence of all intrinsic mental states in terms of their 'aspectual shape,' i.e., intentionality.

How does this relate to the consciousness connection principle? As follows. The causal-dispositional analysis forges a link between all unconscious mental states and those that are conscious by viewing the former as a disposition of the brain to produce the latter. Imagine, for example, a person in a dreamless sleep (cf. *RM*, 159). Though she is unconscious, we may still attribute to this person the belief that snow is white. On Searle's analysis, the unconscious belief in question is the disposition of a particular brain configuration to cause the person to consciously think 'snow is white' — connecting it to consciousness — which is, perhaps, just another way of saying that the mental state in question is 'in principle accessible to consciousness.'

So much for the argument. As a matter of recent history, and unlike the previous arguments concerning universal realizability and the homunculus fallacy, Searle's material on consciousness has already been subject to an extensive and largely critical peer review.[21] So I will close with just two additional observations that need to be made. First, *Searle's claim about the irreducibility of intentionality in (1) does not square with his metaphysics, and is on that account entirely misleading*. He says that intentionality or the 'aspectual feature' cannot be 'characterized solely in terms of third-person, behavioral, or even neurophysiological predicates,' and, moreover, that 'no amount of neurophysiological facts under neurophysiological descriptions constitute aspectual facts' (*RM*, 157-8). But reader beware. Searle is famous for his particular brand of *physicalism*, though he prefers the term 'biological naturalism.' For him intentionality just *is* a neurophysical property of the brain, albeit a high-level one.[22] In an important footnote Searle makes the proper concession: 'For these purposes I am contrasting "neurophysiological" and "mental," but of course on the view I have been expounding throughout this book, the mental is neurophysiological at a higher level' (*RM*, 253, n. 3). Searle is no dualist, after all, and no functionalist either. Hence, if intentionality is indeed a high-level neurophysical property, then intentional descrip-

[21] The Open Peer Commentary to Searle's 'Consciousness, Explanatory Inversion, and Cognitive Science,' *Behavioral and Brain Sciences* **13** (1990) 585-640. The points I make are largely independent of any arguments raised there.

[22] The term 'biological naturalism' was coined by Searle in his *Intentionality: An Essay in the Philosophy of Mind* (Cambridge: Cambridge University Press 1983), 264. According to this view, all mental traits are 'biologically specific characteristics' (*RM*, 90). Echoing an old materialist cry, Searle enjoins us elsewhere to: 'Think of the mind and mental processes as biological phenomena which are as biologically based as growth or digestion or the secretion of bile,' from *Minds, Brains, and Science* (Cambridge, MA: Harvard University Press 1984), 54.

tions *are* neurophysical descriptions — the relevant singular terms denote nothing but neurophysiology and the relevant predicates pick out high-level neurophysical properties — so that (1) is false.

In what sense, then, can (1) be understood to express a truth, at least assuming the framework of biological naturalism? Perhaps this: (1′) intentionality does not reduce to neurophysiology *under other neurophysical descriptions*. Well, no doubt, everything is what it is and not another. Even so, intentionality remains neurophysical in its own right. And if so, there is no need to infer any causal-dispositional analysis from the mere fact that, by (2), 'only' our neurophysiology corresponds to the attributions of unconscious intentional states. Our neurophysiology is fraught with intentionality, or so Searle has always maintained.

Or perhaps Searle means to express this: (1″) intentionality does not reduce to neurophysiology *under the limited descriptive vocabulary of present-day neuroscience, given its predilection for objective, third-person descriptions.* No doubt. Still, notice that both (1′) or (1″) are alike in a crucial respect, being claims about the reduction of facts 'under certain descriptions.' Yet the subsequent proposition (2) is a claim about metaphysics, not vocabulary, about the ontology that lies behind our talk of unconscious intentional states. And, again, it is perfectly consistent with (1′) or (1″) that a rich neurophysiology with high level, biologically natural though intentional properties lies behind all attributions of unconscious mental phenomena. Compare in this regard an objection raised by Clark Glymour, who complains that Searle must equivocate between *facts* and *facts under description.*[23] As Glymour put it, from 'All mental states are aspectual,' and 'No neurological descriptions involve the aspectual,' it does *not* follow that 'No neurological states are aspectual.' Quite so. But if I am right, the inference is much worse, having its minor premise hedged considerably, as (1′) or (1″) will attest.

But all is not lost. Taking high versus low-level facts as our cue,[24] perhaps what Searle means, finally, is this: (1‴) intentionality does not reduce to *low-level neurophysical properties, since all low-level neurophysical*

[23] See Glymour's 'Unconscious Mental Processes,' *Behavioral and Brain Sciences* **13** (1990), 606. The criticism does seem warranted. In his 'Consciousness, Explanatory Inversion, and Cognitive Science,' Searle says that no neurophysical facts have aspectual shape 'under neurophysical descriptions' (587). And he repeats it here (*RM*, 158, 169). Note, too, that a similar problem would arise if Searle means to emphasize the *epistemological* difference reflected in the differing descriptions, i.e., the objectivity of neuroscience versus the subjectivity of consciousness. It may nonetheless be the same facts that are known. Cf. the 'intensionalist fallacy' committed by some Cartesian dualists, nicely discussed in Paul Churchland, 'Reduction, Qualia, and the Direct Introspection of the Brain,' *Journal of Philosophy* **82** (1985) 8-28.

phenomena are nonintentionally specifiable. This proposal has two virtues. First, by distinguishing between high and low-level facts, Searle may locate intentionality within a subject's neurophysiology and thereby preserve his biological naturalism without identifying intentional states with the same neurophysiology that constitutes the ontology of unconscious mental phenomena. Second, the proposal allows Searle to avoid Glymour's charge of equivocation since, by identifying aspectual shape with a high-level neurophysical property, and identifying the ontology of the unconscious with low-level neurophysical facts that have no such properties, then Searle is not talking about the *same* facts under different description when he talks about intentional states on the one hand and the ontology of the unconscious on the other.

Yet even this is unacceptable. For Searle's entire argument must be reinterpreted in light of (1‴), and hence (2) would now read: (2‴) 'the ontology of unconscious mental states, at the time they are unconscious, consists entirely in the existence of purely *low-level* neurophysiological phenomena.' But that should be considered false by all parties in the dispute. Consider the behavior that may accompany *petit mal* seizures — walking, driving, even playing the piano — all extremely complex unconscious behaviors that demand some relatively high-level neurophysical activity.[25] So while there may be some interpretation of Searle's irreducibility claim that both preserves his biological naturalism and generates a plausible argument in the form of (1) through (4), I think we

[24] Talk of theoretical levels is common enough, though different authors imply different things. In the context of cognitive theory, see J.R. Anderson, 'Methodologies for Studying Human Knowledge,' *Behavioral and Brain Sciences* **10** (1987) 467-505.

[25] These cases are mentioned by Searle, who concedes that the behavior is complex, though 'habitual, routine, and memorized' as compared to the behavior that results from consciousness, which has a 'degree of flexibility and creativity' (*RM*, 108). Regardless, the question is the *level* of neurophysiology required for explanation, and it is undoubtedly high for such 'memorized' behavior. Also, while I assume (2‴) is false, even the unrevised (2) is terribly contentious. Why believe that the ontology of unconscious mental states is *exclusively neurophysical*, regardless of level? Cognitivism, for one, postulates a wide range of high-level functionally individuated computational states realized by this neurophysiology that nevertheless operate at an unconscious level, some having intentional properties that cannot be found at lower levels of organization. Remember Marr's three dimensional descriptions mentioned earlier, which purportedly represent visual information from the external world. So the dialectical situation is all too familiar: we still need an argument to show why cognitivism is false, one that does not beg the question at a key premise.

must in all honesty remain skeptical. The consciousness connection principle goes unsupported.

Finally, my last observation concerns the causal-dispositional analysis of unconsciousness expressed in (4). Specifically, what Searle says about the topic forces a fairly innocuous interpretation of the consciousness connection principle that is perfectly consistent with the role of the cognitivist's mental syntax.[26] To see this, note first that the causal-dispositional analysis involves two categories of things. There are unconscious neurophysical states devoid of intentionality, and there are conscious states with their intentional features. Searle sums it up later in this way: 'There are brute, blind neurophysiological processes and there is consciousness, but there is nothing else' (*RM*, 228).

To be as clear as possible, let us construe the relevant states implicated by the causal-dispositional analysis as token event structures (objects having properties at times).[27] They are [x,P,t] and [y,M,t′], where x and y are neurophysical objects, P the nonintentional and unconscious neurophysical property, M the intentional and conscious property, with t and t′ being the times at which the events occur. Moreover, these events must be individuated in such a way as to keep them distinct, since Searle's claim about irreducibility must be that property M does not equal P, which arguably makes the event structures nonidentical; and his causal-dispositional analysis requires that there be two nonidentical events standing (or possibly standing) in the cause and effect relation, with [x,P,t] having the causal disposition to produce [y,M,t′].

Now the crucial point is this: Searle maintains that the nonintentional and unconscious neural event [x,P,t] is a *legitimate mental event* in virtue of the fact that it has the capacity to bring about another event [y,M,t′] that *is* intentional and which *does* enter consciousness. That is what distinguishes neural events that are mental from those that are not. He says this:

[26] Others have argued for the same *conclusion*, but from a different direction by analyzing the connection principle's important proviso that it may be possible 'in principle' for a mental state to fall under the scope of consciousness, and then showing that mental syntax can be accessible in just this sense. See Ned Block, 'Consciousness and Accessibility,' 596; to a lesser extent Noam Chomsky, 'Accessibility "in Principle,"' 600-1; and David Rosenthal, On Being Accessible to Consciousness,' 621-2, all in *Behavioral and Brain Sciences* **13** (1990). Here, on the other hand, our focus is the *causal-dispositional analysis* and what it implies for the connection principle.

[27] For details on the structural view, see Jaegwon Kim, 'Causation, Subsumption, and the Concept of Event,' *Journal of Philosophy* 70 (1973) 217-36; also Lawrence Lombard, *Events: A Metaphysical Study* (London: Routledge & Kegan Paul 1986).

> Of the unconscious neurophysiological processes, some are mental and some are not. The difference between them is not in consciousness, because, by hypothesis, neither is conscious; the difference is that *the mental processes are candidates for consciousness, because they are capable of causing conscious states.* But that's all. All my mental life is lodged in my brain. *But what in my brain is my "mental life"? Just two things: conscious states and those neurophysiological states and processes that — given the right circumstances — are capable of generating conscious states.* (*RM*, 161-2, italics mine)

But not only are certain nonintentional and unconscious neural events legitimately *mental* by reason of their capacity to cause consciousness, Searle believes that this causal capacity makes them *in principle accessible to consciousness.* Speaking of the same neurophysiology, he says in the above quote that 'the mental processes are *candidates for consciousness* because they are capable of causing conscious states.' And in the subsequent passage Searle calls all those neurophysical states which enjoy the designated causal capacity merely 'shallow unconscious' states that are 'in principle accessible to consciousness,' and this, in contradistinction to the allegedly 'deep unconscious' states that are 'inaccessible even in principle' (*RM*, 162). Indeed, remember that the whole point of Searle's argument (1) through (4) is to support his consciousness connection principle, an argument which culminates in the causal-dispositional analysis of unconsciousness.

But here is the rub — if a given nonintentional and unconscious neural event [x,P,t] is both legitimately mental and in principle accessible to consciousness in virtue of the fact that it has the capacity to cause an intentional and conscious event [y,M,t′], then *nothing prevents us from saying that the syntactic states of computational theory are likewise legitimately mental and in principle accessible to consciousness.* Why? Because *they too cause or determine conscious states.* That is, after all, their theoretical role. Any instance of sub-personal level rule following, information processing, inference, or description *via* some mental symbol will earn its keep only insofar as it explains mental states and the resultant behavior of the system in question.

How could Searle have missed such an obvious retort, leaving the door wide open for cognitivism to enter? Perhaps, anticipating his already examined 'Critique of Cognitive Reason,' Searle assumed that the observer relativity of syntax was a foregone conclusion so that syntax simply could not play the role of causing conscious states. But there is another possibility, namely, that Searle wanted more out of his consciousness connection principle than the causal-dispositional analysis would allow. In particular, we cannot ignore the common association between 'consciousness' and 'awareness,' to which Searle pays deference (*RM*, 84); and we cannot deny that computational events are things of which we are typically *unaware*, introspectively speaking. This being so, we have two quite different ideas to consider: being 'connected to consciousness' in the sense that something merely *causes* a conscious state (the innocuous reading sanctioned by the causal-dispositional analysis), and being 'connected to consciousness' in the sense that something *falls within the scope of the subject's awareness* (the more substantive reading aligned with common usage). More precisely, I think Searle vacillates between:

> (a) An unconscious event [x,P,t] satisfies the consciousness connection principle if and only if [x,P,t] has the causal capacity to generate a conscious event [y,M,t′].

versus:

> (b) An unconscious event [x,P,t] satisfies the consciousness connection principle if and only if [x,P,t] has the causal capacity to generate a conscious event [y,M,t′], and "[x,P,t]" is the content of the intentional feature of [y,M,t′].

The point of the additional clause in (b) is to capture the idea that the subject in which these events occur would be *aware* of the cause [x,P,t], and be aware in virtue of the fact that the cause is represented by the intentional feature of the event [y,M,t′] which is by hypothesis a *conscious* event.[28] Putting matters thus, I have previously argued, in effect, that Searle's causal-dispositional analysis of unconscious mental states will only license (a), and that computational events may well satisfy the consciousness connection principle in this sense. On the other hand, computational events would seem to fail by condition (b).[29] Our conscious thoughts are typically *not* about the computational events that cause them. They do not enter consciousness in this way. Can Searle find solace in this fact? Not at all. For by the same token, *neurophysical* events will fail by condition (b), since we are not typically aware of the neurophysical causes of consciousness either. Put more forcefully, given

[28] I have not attempted to explain what would *make* [y,M,t′] a conscious event, only that if it *is* a conscious event, then the cause [x,P,t] could be reflected in the subject's conscious awareness by being represented in its conscious mental episodes. Parenthetically, those who believe that there can be noncausal representation would resist the stipulation in the first clause of (b) that [x,P,t] must have the 'causal capacity' to generate the conscious event. But I am genuinely non-committed, and more than happy to waive the restriction. It will not affect the overall point I am trying to make in the text, viz., that the subject is typically unaware of both computational *and* neurophysical events.

[29] Well, they may not fail by (b) either. For it might be possible 'in principle' for the computational events to fall within the scope of consciousness. See the references in n. 26.

Searle's causal-dispositional analysis whereby the causes of conscious mental states are purely nonintentional, unconscious, neurophysical states, the more substantive reading (b) is not just false but wildly in error — *it implies that our conscious thoughts would be about those brute, blind neurophysical causes!* That point notwithstanding, Searle does, at times, lean towards (b). He says that an unconscious mental state (i.e., the neurophysiology) must be 'a possible content of consciousness' (*RM*, 155), implying that the conscious event [y,M,t'] would, if realized, have the unconscious neural event [x,P,t] as its *content*. Similarly in this passage:

> There are plenty of unconscious mental phenomena, but to the extent that they are genuinely *intentional*, they must in some sense preserve their aspectual shape even when unconscious, but the only sense that we can give to the notion that they preserve their aspectual shape when unconscious is that they are possible contents of consciousness. (*RM*, 160)

It may well be doubted that a given neural event [x,P,t] which is completely devoid of intentionality and consciousness can be accurately described as having its intentionality 'preserved.' Nonetheless, if, as Searle says, it is a possible *content* of consciousness, then the content '[x,P,t]' must be represented in consciousness, exactly in accordance with (b).

My conclusion, then, is twofold. First, if Searle wants a consciousness connection principle like (b), then his supporting argument in terms of the causal-dispositional analysis is a complete *non sequitur.* It does not license anything stronger than (a), where the notion of an event falling within the scope of awareness gives way to the notion of an event merely causing an episode of awareness. Second, and what is worse, Searle's causal-dispositional analysis seems flatly inconsistent with the stronger reading (b), since, again, it would require some conscious awareness of our brute, blind neurophysiology — the only ontological facts within the realm of unconsciousness that Searle assigns to the role of causing conscious mental states.

So let us take stock. We examined Searle's claim about the observer relativity of syntax, starting with his argument concerning universal realizability and later the homunculus fallacy. This, in turn, led to the formulation of an argument concerning symbols and interpreters, all roundly criticized in previous sections. We then cast about for something positive on Searle's behalf, and suggested his consciousness connection principle, which was supported in turn by a series of claims resulting in the causal-dispositional analysis of unconscious mental states. But for the reasons just discussed, this latter argument was also found wanting.

I conclude that Searle has failed to generate any convincing argument against cognitivism.[30]

Received: June, 1993
Revised: June, 1994
Revised: January, 1995

[30] I should like to thank Charles Carr, Thomas Grimes, Terence Horgan, Jaegwon Kim, and also an anonymous referee for their helpful comments on an earlier draft of this paper.