

The semantic challenge to computational neuroscience

*****Draft. Please don't quote.*****

Rick Grush
Department of Philosophy, and:
Center for the Neural Basis of Cognition,
Department of Linguistics,
Department of History and Philosophy of Science
Intelligent Systems Program
University of Pittsburgh

Version: 2.4
Date: 02.26.00
Word count: 7,335

ABSTRACT: I examine one of the conceptual cornerstones of the field known as *computational neuroscience*, especially as articulated in Churchland et al. (1990), an article that is arguably the *locus classicus* of this term and its meaning. The authors of that article try, but I claim ultimately fail, to mark off the enterprise of computational neuroscience as an interdisciplinary approach to understanding the cognitive, information-processing functions of the brain. The failure is a result of the fact that the authors provide no principled means to distinguish the study of neural systems as genuinely computational/information-processing from the study of any complex causal process. I then argue for two things. First, that in order to appropriately mark off computational neuroscience, one must be able to assign a semantics to the states over which an attempt to provide a computational explanation is made. Second, I show that **neither** of the two most popular ways of trying to effect such content assignation -- informational semantics and 'biosemantics' -- can make the required distinction, at least not in a way that a computational neuroscientist should be happy about. The moral of the story as I take it is not a negative one to the effect that computational neuroscience is in principle incapable of doing what it wants to do. Rather, it is to point out some work that remains to be done.

1. Some distinctions.

In discussing the relationship between the brain and computation, it will be helpful to begin with the following crucial distinction:

A. *Computation (or more felicitously a computer) as a tool for simulation.* Computers running the right software can be used to simulate a vast range of phenomena, from the antics of sub-atomic particles within a nucleus to the ways in which minor variations in conditions just following big bang might effect the distribution of matter throughout the universe. One can simulate economic systems, human psychological performance, weather systems, and even computational systems themselves. What is important to note is that one can design and run computer simulations of a system or phenomenon without in any way taking a stand on whether or not that system or phenomenon is itself computing anything.

B. *Computation as a theoretical stance in cognitive neuroscience.* Many neuroscientists believe that one can shed light on the operation of biological neural systems by treating them as themselves carrying out computations. That is, unlike planets in the solar system, which *act in accord* with various functions but are not themselves computing those functions, the idea is that neural systems *really are computing* this or that function, *not merely acting in accord with it.*

With respect to B, we can make two further distinctions, depending on what one means by ‘compute’ or ‘computational’.

B1. *The brain (or parts thereof) implements some specific general-purpose computational architecture* (e.g. a (finite-tape) Turing machine, or some other finite-state automaton, for example). Few neuroscientists take this position at all seriously. Nobody really thinks the brain stores symbols corresponding to discrete 1s and 0s and operates on them in a serial manner according to a state-transition table, etc. Nonetheless, it is useful to bring this position to center stage if only to be sure to avoid it.

B2. *The brain (or parts thereof) computes in the broad sense of implementing computable functions.* The presumption is that it is that by implementing such functions the brain processes information, and thus we might call this view the view that the brain is a *computational information processor* (more will be said about information processing below). This is the serious contender, and most neuroscientist, especially those who call themselves computational neuroscientists, readily subscribe to it. In this broad sense, (implemented) Turing machines, finite- and combinatorial-state automata, including most (maybe all) connectionist networks, and many dynamical systems will count as things that compute.

In the remainder of this paper, I will simply take it as understood that all mention of computation is to be understood in the broad sense, and therefore that *computational neuroscience* is the field of study defined by the guiding assumption that biological neural systems are best understood as systems that process information by implementing computable functions.

This seems innocent and uncontroversial enough, but I will try show that the innocence is quickly lost. The trouble will come when one tries to keep B2 and A distinct. That is, when one tries to keep the notion of being a computational information processor i) broad enough so that we can count things other than digital computers as computational information processors, while at the same time keeping it ii) narrow enough so that not any system that can be simulated will count as a computational information processor.

That all is potentially not well can be seen from the following passage from Churchland, Koch and Sejnowski (1990), an article that is considered one of (perhaps the first and best of) the defining expressions of the project of computational neuroscience:

A physical system is considered a computer when its states can be taken as representing states of some other system; that is, so long as someone sees an interpretation of its states in terms of a given algorithm. Thus a central feature of this characterization is that whether something is a computer has an interest-relative component, in the sense that it depends on whether someone has an interest in the device's abstract properties and in interpreting its states as representing states of something else.

The first thing to note is the ambiguous use that the term 'represent' gets in this passage. Phrases such as "representing states of some other system" suggest that what is crucial is some semantic relationship between a physical state of the brain and some other external object or state of affairs in the environment of the creature whose brain is under discussion. Let us call this sort of interpretation an environmental-semantic, or simply *e-semantic* interpretation. On the other hand "interpretation of its states in terms of a given algorithm" suggests that the crucial semantic relationship is between a physical state of the brain and a variable, or value of a variable, of an abstract algorithm. Let us call this sort of interpretation an algorithm-semantic, or simply *a-semantic* interpretation. One can't help but notice the easy slide from e-semantics to a-semantics in the first sentence of the quote, for example -- and then back again in the last sentence. Though e-semantics is seldom far from the surface of Churchland et al.'s discussion, by and large it is a-semantics that seems to be central. This gets spelled out just further on:

In a most general sense, we can consider a physical system as a computational system just in case there is an appropriate (revealing) mapping between some algorithm and associated physical variables. More exactly, a physical system computes a function $f(x)$ when there is (1) a mapping between the system's physical inputs and x , (2) a mapping between the system's physical outputs and y , such that (3) $f(x) = y$

This is clearly a-semantics, and it is difficult not to find this a bit unsettling. According to this proposal, everything is computing the function specified by equations that govern its physical behavior. Notice that exactly this constraint -- the existence of a (revealing?) mapping between the physical states of some system and some algorithm (or computable function) -- is both necessary and sufficient for being able to simulate that physical system by means of a

computer.¹ That is, on the a-semantic account of what makes a physical system a computer, an account that is taken from the canonical defining expression of computational neuroscience, B2 collapses to A without remainder.

Perhaps this is what drove Churchland et al. to make the concession regarding the interest-relativity of computing. But these concessions don't solve the problem, they merely highlight it - for they imply that the only difference between A and B2 is that the B2 cases are the A cases such that someone cares enough about them to actually construct a computer simulation of them. We wanted to know when some physical thing, like the brain, really is a computer -- really is computing. Churchland et al. seem to have simply given up at just the moment we needed their help most.

But I don't like to give up unless it is absolutely necessary. Perhaps we can get somewhere by making a further distinction, a distinction blurred in the expression of B2, between two senses in which B2 can be understood, a distinction that hinges on whether we are doing a-semantics or e-semantics:

B2a: The brain (or parts thereof) computes in the sense that it implements computable mathematical functions. This is what we have just been discussing. On this view, a

¹ That it is necessary can be seen by noting that implementing algorithms is just what computers do. If there is no algorithm, then one cannot implement it on a computer. Of course, one may not know the exact algorithm before one does the simulation. One might find out what the algorithm is by training a connectionist network, for instance. The sufficiency follows by definition. Given that computers are general purpose (one may need to make idealizations about amount of memory here, of course), meaning that they can execute any computable function, having a computable function is sufficient for being able to implement it in a general purpose computer.

system is a computer if there is a revealing mapping between its states and the input, output, and any 'intermediate' variables of some computable function. The qualifier 'revealing' need, and perhaps can, mean no more than that simulations exploiting that algorithm shed some light, for someone, on the operation of the system. This much will be true of any successful computer simulation.

B2e: *The brain (or parts thereof) computes in the sense that it processes information -- it deals with what genuinely are information-carrying states -- e.g. states that carry information about objects or states of affairs in the environment.* On this view, it is e-semantics that is doing the work.

Notice that on the surely plausible assumption that the processing of information occurs according to tractable law-like causal processes, the B2e cases are a subset of the B2a cases. That is, all B2e's will also be B2a's, since any tractable process is a B2a. The hope is that the same considerations that serve to differentiate the B2e's from the other B2a's will also serve to differentiate the systems that really are computing a function, from those that are acting in accord with, but not computing, a function. In the next two sections I will briefly present two case studies of computational neuroscience in action with an eye towards clarifying the distinction between B2a and B2e. The examples are a bit dated (each from the late 80s), but that is not relevant. They have been chosen precisely because they have been taken to be parade-case examples of successful computational neuroscience, and because they bring out the difference between B2a and B2e.

2. Computational Neuroscience I: Koch

First an example at the level of the single neuron. Christof Koch (1990) reports studies, done in collaboration with the Paul Adams laboratory, on the bullfrog sympathetic ganglion cell:

We chose as the object of our study type "B" bullfrog sympathetic ganglion cells ... for which the description of the various macroscopic currents is fairly complete. Due to their lack of dendrites, synapses are formed on or near the cell body and space clamp problems are absent. The aims of this work are (1) to separate the total ionic current into various distinct components, (2) to develop empirical equations that approximately describe the behavior of the currents under physiological conditions, and (3) to compare the computer simulations with cell responses during various experimental paradigms, for instance using pharmacological blockers.

The equation used to model the neuron is

$$(1) \quad I = g_{Na}x_{Na}^2y_{Na}(V - E_{Na}) + g_{Ca}x_{Ca}y_{Ca}(V - E_{Ca}) + (g_Ax_Ay_A + g_Mx_M + g_Kx_K^2y_K + g_Cx_C + g_{AHP}x_{AHP}^2)(V - E_K) + g_{LEAK}(V - E_{leak}) + cdV/dt$$

What each of these variable means is not to the point. What is to the point is that this equation is an equation describing the physical processes that govern the relevant electrical behavior of the neuron. The equation is an abstract mathematical object, and the variable names have been chosen so as to suggest the names of the physical parameters to which they map (where 'map' is clearly being used here in an a-semantic sense). For example, the variable g_{LEAK} is the a-semantic interpretation of membrane leak conductance.

The point of the investigations on which Koch reports is to determine whether or not equation (1) does in fact accurately describe the physical behavior of the neuron. This is tested by simulation studies, in which the behavior of a virtual neuron (whose behavior is guaranteed to match (1) because (1) is used to model it) is compared to the behavior of a real neuron under a

variety of matched real/simulated conditions, including the influence of (real and simulated) pharmacological blockers.

The result of the simulations, of course, was that the behavior of the simulated neuron matched (to within some acceptable degree of accuracy) that of the real neuron in a variety of conditions, the match being close enough to merit some confidence that (1) was in fact the equation describing the physical behavior of that system.

This is a clear case of B2a. The only semantic relation appealed to is between states of the neuron and variables of an equation. There is no pretense to the effect that *information processing* is going on in the neuron, and hence there is no pretense to e-semantics. The neuron satisfies the Churchland et al. a-semantic requirement for performing a computation. But this is no surprise, since *any* tractable physical process satisfies this requirement. Of course, it may be the case that the neuron does process information because of the way it behaves, but if so, this fact is entirely external to and incidental to this simulation.

3. Computational Neuroscience II: Zipser and Andersen.

One of the most fascinating and successful examples of computational neuroscience is the Zipser and Andersen model of the response properties of (some) neurons in posterior parietal area 7a.

First the initial biological data. Lesion data had long suggested that this cortical area is crucial for spatial representation, but little was known about how this was accomplished. Single cell recordings in the early and mid 80s (e.g. Andersen et al. 1987) showed that there are at least 3 groups of cells in this area: some have responses tied to the retinal location of the projection of a visually presented stimulus; a second group responds to the orientation of the eye in the head; and a third group has responses that are influenced both by the location of a visual stimulus on the retina (retinal location), and the orbital orientation of the eye. Such cells fire maximally only given a certain combination of retinal location and orbital position.

Next the simulation. Zipser and Andersen designed a connectionist network whose inputs were i) the location on the retina of a given stimulation, and ii) eye orientation of a simplified two-dimensional eye. The network was trained to learn the direction of the stimulus relative to the 'head' given these two pieces of information. The result was that the model learned the task, meaning that it learned to provide the correct output given the two inputs. Furthermore, it was discovered that the response profiles of 'hidden' units of the network had response properties that seemed to be similar to the response properties of actual neurons in area 7a.

The payoff, however, was that the connectionist network could be analyzed in great detail to determine exactly how it was solving the problem. It was learned that the hidden units were implementing planar gain fields, in which each unit had a gaussian retinal receptive field, but in addition had its activity modulated linearly by the orientation of the eye. In effect, a given unit would fire most strongly with a given combination of retinal stimulus location and eye

orientation. (It was also determined that the same mechanisms could be used by a connectionist model to compute body-centered locations, by having hidden units that acted as planar gain fields combining retinal location, eye position, and head orientation. Subsequent single cell recordings confirmed the existence of cells with such response profiles.)

The situation with this simulation is more complex than the situation with the last. In this case, as in the last, there is an a-semantic relation between physical states of neurons and variables of a mathematical function. That is, we can suppose that the function computed by an idealized posterior parietal neuron is something like $f = (\vartheta - \vartheta_p) g(r - r_0)$, where f is the a-semantic interpretation of firing frequency, $\vartheta - \vartheta_p$ is the interpretation of firing frequency of one pre-synaptic neuron, and $g(r - r_0)$ is the interpretation of firing frequency of the other pre-synaptic neuron. This is a B2a explanation: The state of the cell is some function of the states of the cells that synapse on it.

But in addition to this a-semantic mapping, there is an e-semantic mapping: f is the e-semantic interpretation of stimulus distance from preferred direction relative to the head (high firing rate means low distance), $\vartheta - \vartheta_p$ is the interpretation of the difference between the actual and preferred eye orientation, and $g(r - r_0)$ is the interpretation of (a gaussian of) the distance from the retinal location of stimulation from the receptive field. This provides for a B2e explanation: The cell is processing information about retinal location and eye orientation to provide information about direction of stimulus relative to the head.

4. Representation.

We have before us two examples of computational neuroscience in action. The study by Koch involving the bullfrog ganglion cell appears to involve no more than would be involved in any example of computer-simulation-cum-experimental-testing endeavor, of the sort familiar in computational physics, economics, meteorology, and dozens of other areas. This is by no means to say that it is uninteresting. Quite the opposite, such interaction of experiment and computer simulation has proven to be an incredibly powerful tool for shedding light on the operation of complex systems whose principles would otherwise remain hidden. The point is that this model is a clear case of B2a, and B2a is just a tarted-up version of A.

The Zipser and Andersen model (and subsequent experiments based on it), however, seems to be a good example of a B2e. It sheds light on how a set of neurons interact so as **to represent directions of stimuli in egocentric space**. Of course, one *can* describe the behavior of the neurons in the posterior parietal in a thin B2a manner, as discussed in the previous section. But one also has the B2e description available. One plausible suggestion is that it is the availability of such an alternative, *representational*, description that sets B2e cases apart from B2a cases. This is of course what one would expect, since B2e cases just are cases where there is a semantic relation established between the neural states and something in the environment. This something

will be what the neural state represents. (The question what it is that makes this representational description available and appropriate -- what establishes the semantic relation -- will be taken up shortly.)

So, if there were some principled means to determine which states are representing aspects of the environment, we could exploit this to determine which of the B2a cases are in fact B2e cases. We would have the means to distinguish those systems that are genuinely computational in the required sense, and there would be no danger of computational neuroscience being assimilated without residue into the general category of computer simulation studies.²

I will assume that a bald interpretationalist stance here is unacceptable. That is, I am assuming -- as should any self-respecting computational neuroscientist -- that there is a fact of the matter concerning whether or not a given neural state is representing something: a fact that is neither created nor imperiled by the interpretive whims of whomever might be examining the neural system. But the conviction to the effect that there is such a fact of the matter is not enough. What we need is some means to explain why it is that *this* state does, but *that* state does not, really represent something. If we lack such means, then computational neuroscience will have to proceed on faith and in ignorance.

Now I, like many people who spend time worrying about issues of representational content, happen to think that such means can be provided. However, the theories about how this happens

that have the most currency are, I think, unworkable. This is not the time to go into a rather involved positive account of my own (for this, see Grush, in preparation). But this is the time to show why neither of the two most popular and initially plausible accounts will work.

5. Informational Semantics

A groundbreaking but ultimately untenable attempt to show how some things are intrinsically representational is the informational semantic approach, whose *locus classicus* is Dretske (1981).³ According to this approach, one can appeal to the information carried by the state of some system in order to fix the content carried by that system. For instance, suppose that a given neuron in my cortex fires at a rate faster than 50 Hz when and only when there is a square in my visual field. If this is the case, then one can say that that neuron's state of firing over 50 Hz carries information to the effect that there is a square stimulus in my visual field (and also: the neuron's state of firing *under* 50 Hz carries information to the effect that there is *not* a square stimulus in the visual field). Anyone familiar with the neurophysiology literature on perception, cognitive mapping, motor control, etc., will recognize that this approach is widely assumed by neuroscientists, who look for causal covariation to establish that something's representational content (a moving bar, a face, a twitching paw, a 'place', etc.).

² Though I arrive at them by different means, my conclusions here are similar to those of Shagrir (199@).

³ In what follows, I will be providing only the roughest sketch of a proposal that has had many clever defenders who have provided responses to the sorts of concerns I raise. I am not here trying to kill informational semantics, but merely recounting, in caricature form, some of the main sources of pressure that led to its death.

The problem with this suggestion is that information is as ubiquitous as are law-governed causal processes. The temperature of the outer surface of my coffee cup is a well-behaved function of the following: original temperature of the coffee, amount of coffee, original temperature of the cup, time since the coffee was poured, the ambient temperature, and heat-transfer properties of the cup. Given that all but the room's ambient temperature is fixed when the coffee is poured, the cup is processing information about the ambient temperature of the room. In general, any physical state of anything carries information about the states of other physical entities that enter into causal dialogue with that state, and hence it will be possible to view any causally fertile (law governed) process involving the evolution of physical states *of anything* as information carrying. (Some examples of *ersatz* information carrying in the brain are provided in the next section.)

As it happens, problems such as these, and many others have led to the collapse of informational semantics, its prevalent role in the pre-theoretical intuitions of neuroscientists notwithstanding. Evidently we must look elsewhere for our semantic salvation.

6. Biosemantics.

Yet another proposal, one that is increasingly popular, is appeal to biological functions (see Millikan 1984 for the *locus classicus*; see also Millikan 1989). On this proposal, the proper function of *d* in *S* is to detect *c* iff *d* was replicated because it carried information about *c*, and for

our purposes we can take 'detect' to have the force of 'represent'.⁴ In short, the idea is that the cell in my posterior parietal cortex (PPC) represents the direction of a visual stimulus in virtue of the fact that that PPC neuron (or more plausibly the architecture of the PPC which induces such neurons) was replicated because it carried information about direction of a stimulus relative to the head.

How does this solve the problem that bedeviled informational semantics? Consider the following example. The skin in the bottom of my feet has cells that respond to pressure. Given this, and other things being equal, the more I eat, the more these cells will fire, since the more I eat, the more I weigh, and the more I weigh, the more pressure on the bottom of my feet. Furthermore, there are cells in my primary somatosensory cortex that are responsive to the activity of these pressure-sensitive cells. So in terms of carrying information, these cells in my somatosensory cortex carry information about how much I have eaten recently. This is just another example of the ubiquity of information.

But on the biosemantic account, those cells are not representing how much I have eaten recently, because the response profiles of these cells was *not* replicated because they carried information about my eating behavior. The explanation of their replication would rely on other considerations altogether, having to do in part with detecting features of the surface with which the feet are in contact. There are other sets of neural structures that were replicated because the information they carried was useful for -- *and actually used for* -- the regulation eating.

So the biosemantic gambit is this: with a prior notion of some biological function in mind, such as facilitating accurate motor control, one can view certain physical states of the organism as contributing to this function because they carry information that is crucial for the appropriate

⁴ As was the case with informational semantics, my gloss here ignores a number of central refinements and

execution of that function. Information about where something is is crucial for my ability to grasp it, for example, whereas the information in the soles of my feet about how much I ate for dinner, while genuinely informational, is not serving any function (though it could be incorporated into a system that does -- imagine a creature whose foraging behaviors were triggered by the pressure sensed on the bottom of its feet, so that as the creature became lighter, it had a greater drive to find and eat food). Complex biological systems have components which are such that an explanation of why those structures are the way they are appeals to the fact that their carrying certain sorts of information facilitated their evolutionary selection. This will be true of only a small subset of such information-passing states.

This sort of account has wide and expanding appeal within the philosophy of psychology literature on content. But I will now show that its appeal notwithstanding, it is not a proposal that the defender of computational neuroscience can embrace without paying a rather high price.

Physical systems, *qua* systems describable by the laws of physics, are state-determined systems. This means that the future behavior of the system is determined (perhaps statistically determined) by its current state. If a ball is at location x_1 with momentum y at t_1 , then in absence of any forces acting on it, it will be at location x_2 at t_2 . What is crucial to note is that the laws of physics *don't care* how the ball got to location x_1 , or how it came to have momentum y . One will not find different laws depending on whether the ball was struck by another ball, or was pushed by a hand or magnetic field, or expelled a billion years prior to t_1 from a nova. How it got into

features of Millikan's proposal, but those refinements and features are inessential to my purpose.

that state is irrelevant. All that is relevant to its behavior -- all one needs in order to apply the laws of physics and determine what the ball will do -- is its *current* state and (the forces currently acting on it, if any).

If we leave the realm of physics momentarily, we can see that it is often the case that things other than an object's narrowly physical state can be crucial to its identity and status. For example, whether a dollar bill is genuine or counterfeit is not determined by its physical state. It is determined by who printed it and where. Two pieces of paper might be in identical physical states -- molecule-for-molecule identical -- and yet it can nevertheless be true that one is a genuine dollar bill and the other is not. Now we can always decide to include anything we want into a description of something's state. That is, we can expand our notion of a *state* to include more information about it than is required for application of the laws of physics. We might, for example, decide to include a new variable in the state of all objects that indicates their place of origin -- perhaps a variable *T* that takes a binary +/-, and an object is *T+* if it was made by the US Treasury, and *T-* otherwise.

If we expand our conception of a thing's current state in this way, then we can make it the case that whether or not a piece of paper is a genuine dollar bill can be determined by its current state. And given that locations and processes of paper production and printing are physical things and processes, we might not feel *too* uncomfortable in continuing to call this expanded state description a physical state description. But this should not divert attention from the fact that an object's *T*-value is unlike its mass, momentum and position in a number of important respects.

For example, I can take a piece of paper into a lab and determine its mass, momentum, etc. But if I want to determine its T-value, the physicist and chemist are useless. I will need to hire a private investigator, perhaps, to try to track the history and origin of the piece of paper.

Now on the biosemantic account, whether or not a state is a representation is analogous to the question whether or not a piece of paper is a genuine dollar bill. It is a function of the history of those mechanisms that support that state, and not a function of the current (narrowly) physical state of the brain. To illustrate: suppose that tomorrow a bolt of lightning strikes a swamp, and the resultant release of energy combined with the chemical ingredients at hand produces an exact molecule-for-molecule identical replica of you, call it *swamp-you* (let's suppose for the sake of example that just like you, swamp-you is created with a book in hand, and is looking at a page just like the one you are currently looking at). The odds of this actually happening are entirely irrelevant.

Now let us suppose that you are looking at the word 'swamp' at exactly the time when swamp-you is created, and in fact swamp-you's eyes are also directed at an inscription of the word 'swamp'. Now if biosemantics is correct, then some cell in your cortex is firing because it is representing a line with a certain orientation, and this is part of what allows you to recognize the word 'swamp'. However, according to the biosemantic proposal, swamp-you, although being in a molecule-for-molecule identical state to you when you are perceiving the word 'swamp', is not perceiving anything, not representing anything, not thinking anything. Swamp-you's 'head' quickly swivels around, and produces a sound that we would categorize as an utterance of the

form 'How the hell did I get in a swamp?' (or whatever you would produce if you suddenly found yourself in a swamp while in the middle of reading a book). But, according to the biosemantic proposal, appearances notwithstanding, swamp-you is not thinking anything, not perceiving anything, because not representing anything.

Why? For the simple reason that in such a case the explanation of why a certain cell in swamp-you fires in a certain way would make *no appeal to evolutionary pressures*, as this swamp-you has *no evolutionary history*. Be careful not to be taken in by the fact that swamp-you is just like something that *does* have an evolutionary history. This is not relevant. (Being *just like* something that is worth \$20 doesn't make counterfeit \$20s really worth \$20.) The explanation of why such and such a potential is present on this membrane (i.e. why this cell is firing) would appeal to physical events immediately preceding the presence of the potential, and an explanation of why the physical system was structured so as to have that sort of effect would appeal to the contents and configuration of the goo in the swamp before the lightning strike, and the details of what happened upon the strike. Given this, it follows that on the biosemantic account swamp-you is not really representing anything, since on the biosemantic account something is a representation only in virtue of being the product of the correct evolutionary selection-pressures. And on the other hand, the explanation of why you are representing such things is not given by the details of what is happening in your brain.

Many people will take this to be a *reductio* of the evolutionary-teleosemantic account, as do I. Nevertheless, the counter-intuitive conclusion has been embraced by proponents of this account -

- an admirable example of bullet-biting at the very least. My goal for now is not to provide any (more) reason to abandon biosemantics, but rather to point out that whether or not biosemantics is correct, its adoption provides no comfort to the computational neuroscientist.

Simply put, evolutionary history is not something that shows up in the occurrent causal operation of the brain, any more that a dollar bill's T-value shows up in its occurrent causal dealings. Note that all the computational-neuroscience in Heaven and Earth would fail to detect a difference between swamp-you's brain and yours -- they are, after all, molecule-for-molecule identical. Yet on the biosemantic account, your brain, but not swamp-you's, is really computing/representing; you, but not swamp-you, have a mind. So if biosemantics is correct, then the following questions are *not* among those that computational neuroscience can answer *or even address*: Are cells in your visual cortex representing things in the environment?; Are cells in your hippocampus representing locations in allocentric space?; What is the relation between the brain and the mind?; etc. You get the idea. In effect, every question that cognitive neuroscience ought to be able to answer if in fact it is the physical structure of the brain that explains thought, cognition, imagination, perception, etc., will turn out to be a question beyond the reach of cognitive neuroscience and computational neuroscience. If biosemantics is correct, we need to consult not the neuroscientist in order to determine if the cell in your brain is representing a face or a line or a color; rather, we hire a private investigator.

There are now two objections to what I have just said that I want to address. The first is that the conclusion I draw isn't as bad as I make it out to be, it just means that we need to be

interdisciplinary in our approach to cognition by including some evolutionary biology. The second is that examples like swamp-you are irrelevant to actual science, their ability to fascinate metaphysically inclined philosophers notwithstanding.

***Objection One:** OK, fine. So neuroscience by itself can't tell us which things are really representing, and hence can't tell us which things are really computational. So we enlist the aid of the evolutionary biologist. And we still have a perfectly legit scientific enterprise. What's the big stink?*

Reply One: If it were this easy, then there would be no stink. But it's not this easy. First point. This is not the sort of innocent appeal to interdisciplinary study that is rather popular in cognitive science. The innocent variety is motivated because all current windows on the brain and cognition suffer from technical limitations -- we can't do single cell recordings of every cell in a human brain while it is doing math, for example. And since each technique (single cell recordings, PET, ERP, evolutionary biology, etc.) supplies some constraints, we will do well to take advantage of those constraints. But the biosemantic appeal to evolutionary biology is different. *It would remain even if we knew all the occurrent physical/causal things there were to know about the brain in question.* Any self-respecting physicalist should find this an unacceptable pill to swallow.

Second point. Evolutionary biology is of exactly no help here anyway, because it begs all the crucial questions. Specifically, the evolutionary biologist is not in the business of taking some animal, and performing experiments on it, or observing it, or whatever, in order to tell whether it

is i) a real product of biological evolution, or ii) a molecule-for-molecule replica of something that perhaps was: a swamp-animal, so to speak. The evolutionary biologist simply assumes that there are no swamp-animals running about, and with that reasonable but unverified assumption in hand proceeds to determine what the most likely lines of evolution, series of adaptations, etc., were in the history of the species of which this animal is, *presumably*, a member.

Objection Two: This swamp-man mumbo-jumbo is fine for metaphysically inclined philosophers, but neither philosophers who understand science, nor scientists themselves, should take it seriously. Such a creature would be so improbably that we need not take its possibility seriously. And even if one were to appear, we would have a misconception about that one particular individual, but this would leave the rest of our science completely unaffected.

Reply Two. This objection simply fails to recognize the nature of the problem that swamp-you creates. The problem is that if we expect neuroscientific explanations of things like consciousness, representation, psychology, and the rest to be forthcoming, then swamp-people - - molecule-for-molecule (and hence neuron-for-neuron) replicas *without* a psychology, consciousness, or representations -- should not be possible *at all*. This is entirely parallel to saying that if special relativity is true, then massive particles being accelerated past the speed of light is not possible *at all*. I take it that the following line of reasoning would be scientifically disreputable: "Jones has shown convincingly that if we were to appropriately apply 10^{80} joules of energy to this dime, it would have speed $c + s$ (for some positive finite speed s). But this should not worry those of us who believe in relativity. The odds of anything ever having that much energy directed to its motion is almost nil, and in point of fact on our planet this simply

never occurs. Perhaps wild-eyed philosophers should worry, but no real physicists should take this at all seriously." I take it that this would not be an intellectually reputable thing for a scientist, or anyone for that matter, to say. It might be reputable to deny the antecedent, claiming or arguing that in fact 10^{80} joules *would not* accelerate anything past c . That would be fine. What is NOT fine is to accept the antecedent, accept that 10^{80} joules would accelerate it past c , and then act as though its low likelihood of occurring, or the fact that around here it simply doesn't in fact occur, or that even if it did occur it would just mean that our theory got *one* out of billions of force-acceleration explanations wrong, somehow renders this fact unproblematic to the acceptability of our apparatus physical explanation: explanatory apparatus that has relativity as a central core.

In *exactly* the same way, it is simply intellectually disreputable to maintain both biosemantics (which entails that swamp-people are possible) *and* the claim that neuroscience will be able to explain consciousness, representation, psychology. Now of course, it seems to me that there is an obvious and easy option that presents itself here: ***ditch biosemantics, pronto***. With all apologies to Millikan and company, I think we should bronze and polish it, and stick it in the museum of brilliant, noble, admirable, yet ultimately wrong theories that the history of philosophy boasts.⁵ Then we can accept the reasonable premise that in the extremely unlikely event that a lightning strike were to create a molecule for molecule replica of you, it would represent and be conscious, just like you. And we can also accept the idea that explanations for

⁵ I hasten to admit that, aside from pointing to the counter-intuitive example of swamp-people, I have provided no *argument* here for rejecting biosemantics. I am making a recommendation. But I expect it to be

cognitive phenomena will be, eventually, produced by cognitive neuroscience. The price that neuroscience pays for ditching biosemantics is that in doing so, it ditches one attempt to solve the problem of determining which things really are computational and which aren't. But neuroscience was getting the short end of that stick anyway. It solved the problem only by denying neuroscience any say in the matter.

7. Conclusion.

The thrust of this article has been negative. I have tried to show that given the currently available theoretical tools, one cannot mark out computational neuroscience as an enterprise whose job is to explain the neural basis of cognition. What is missing is the ability of computational neuroscience to provide an explanation, a theory-internal explanation, of why some neural states are representations, and what the semantic import of those neural states that are representations is. Informational semantics fails to distinguish any useful subclass of states as representations, and biosemantics (as with any externalist content assignment scheme) even if it works, does so only by depriving computational neuroscience *per se* of any claim to be able to explain cognition. If my analysis is correct, what is needed is an internalist semantics -- a theory that can explain, given only appeal to mechanisms internal to the brain itself, why some of those states have

convincing only upon the provision of some arguments, and preferably a good alternate account. The alternative account will be along in Grush (in preparation),

semantic import, and what that import is. I think this can be done, but my defense of this claim will have to wait for another day.⁶

Acknowledgements:

I would like to thank Peter Machamer and Gualtiero Piccinini for feedback on an earlier version of this paper, and participants of the Fifth Pitt-Konstanz Colloquium in Philosophy of Science (Konstanz, Summer 1999) for discussion of a still earlier version of this material in talk form.

References:

Andersen, R.A., R.M. Siegal and G.K. Essick (1987). *Experimental Brain Research* 67:316-322.

Churchland, Patricia S., Christof Koch, and Terrence Sejnowski (1990) What is computational neuroscience? In: Schwartz, Eric L. (ed. 1990). *Computational Neuroscience*. Cambridge, MA: MIT Press.

Dretske, Fred (1981). *Knowledge and the Flow of Information*. @@@

Grush, Rick (1997) The architecture of representation. *Philosophical Psychology* 10(1):5-23.

Grush, Rick (in preparation). *The Machinery of Mindedness*.

Koch, Christof (1990). Biophysics of Computation: Toward the Mechanisms Underlying Information Processing in Single Neurons. In: Schwartz, Eric L. (ed. 1990). *Computational Neuroscience*. Cambridge, MA: MIT Press.

Millikan, Ruth (1984). *Language, thought, and other biological categories*. Cambridge, MA: MIT Press.

Millikan, Ruth (1989). Biosemantics. *Journal of Philosophy*, 86(6):281-297.

Shagrir, Oron (199@). @@@. Unpublished PhD dissertation, University of California, San Diego.

⁶ See Grush (in preparation) *The Machinery of Mindedness*.

Zipser, David, and Richard A. Andersen (1988). A back-propagation programmed network that simulates response properties of a subset of posterior parietal neurons. *Nature* 331(6158):679-684.