Bruce F. Katz

Fixing Functionalism

Abstract: *Functionalism, which views consciousness as the product of the processing of stimuli by the brain, is perhaps the dominant view among researchers in the cognitive sciences and associated fields. However, as a workable scientific model of consciousness, it has been marred by a singular lack of tangible success, except at the broadest levels of explanation. This paper argues that this is not an accident, and that in its standard construal it is simply too unwieldy to assume the burden of full-fledged theory. In its place, a reduced functionalism is introduced by applying the principle of parsimony successively to the elements of standard functionalism until only a minimal framework remains. This simpler account states that consciousness is a function of instantaneous causal relations between processing elements rather the putative algorithm such relations are instantiating. It is then argued as a corollary that the only such relations that matter are those in which reciprocal influences are at play. Thus, purely afferent and efferent causal relations are pruned from consideration. The theory resulting from the addition of this corollary is shown to have good correspondence with a number of recent neurophysiologically-motivated approaches to consciousness, including those that stress the importance of reentry, those that view synchrony as a key independent variable, and those that highlight the importance of the accessibility of conscious contents to multiple processing modules. In addition, the theory is shown to be consistent with recent results in the literature on masking, and those in the literature on binocular rivalry. The paper concludes by arguing that the theoretical and empirical difficulties inherent in consciousness research imply that the principle of parsimony must occupy a more central role in consciousness research than it would in ordinary scientific discourse.*

Correspondence:
Bruce F. Katz, Department of Electrical and Computer Engineering, Drexel University. *Email: katz@cbis.ece.drexel.edu*

Journal of Consciousness Studies, **15**, No. ??, 2008, pp. ??–??

Introduction: The Construction of Consciousness

This paper will argue that the traditional construal of functionalism is too complex to be a workable scientific theory of consciousness, and introduce an alternative, simpler theory in its place. The argument begins by introducing the notion of a constructor, which is intended to describe just how the mental supervenes on the physical. Chalmers (1996a) has argued persuasively that although it is likely that consciousness nomologically supervenes on the physical, it does not do so logically. Nomological supervenience implies that the universe, as a matter of scientific law, would imbue an identical atom by atom copy of a person not only with identical behavior, but also an identical inner life. Logical supervenience means that there is also a necessary connection between physical processes and consciousness in addition to this law-governed connection. The failure of logical supervenience follows from the fact that it is conceivable that the identical copy is completely lacking in an inner life. There is simply no entailment from physical facts to mental facts in the way, say, that once given all the physical facts about snow one also must conclude that it is white, it can be skied upon, etc. Given all the facts about the brain, one would have no reason to believe that it generates consciousness if one where not conscious oneself. Kripke (1972) has put it more evocatively: After God created matter and energy and the laws that these obey, he still had more work to do, namely, the creation of laws dictating how the mental follows from the physical.

Let us make the nomological connection between the mental and the physical explicit by identifying this transformation with the symbol C (or in more descriptive terms, the constructor). C may be defined as the transformation that governs the supervenience of the mental on the physical. That C is a function in a mathematical sense follows from the definition of supervenience, namely, that for x to supervene on y means that different x's imply different y's. Hence, C can be conceived of as a mapping from the physical to the mental such that it is impossible for two identical brain states to result in different mental states.

It is easy to show by example that the derivation of C is non-trivial, if that is not already obvious. Consider the case of binocular rivalry, illustrated in Figure 1. Separate images are presented to the two eyes. Under normal conditions, these images would be slightly offset and 'fused' by the brain to yield the perception of depth. However, if the images are radically different, as in the example illustrated, a competition occurs between the images and only one is perceived at any one

time (Blake, 2001). In this case, the eliminativist W.V.Quine is not seen and the genie is. The role of C in this example would be to take the brain state or processes associated with the presentation of these stimuli and return either one or the other image. What makes the operation of C non-trivial in the case of rivalry is that superficially, at least, there appears to be no single process or brain region that is unequivocally correlated with the seen image; we will return to this topic toward the end of this paper.

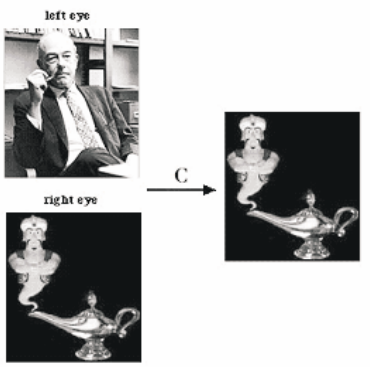

Figure 1. When two conflicting images are presented to the eyes, only one of the images at any given time will be consciously perceived.

What is assumed at the start, however, is that C works by looking at how the brain processes its inputs. That is, C will be associated with a form of functionalism, rather than a purely physicalist account of mind. Functionalism is the theory, or more accurately, the proto-theory that states that the mental arises because of the causal role that the brain plays in transforming input to outputs. Functionalism thus identifies consciousness not with a particular physical platform but with the means by which a particular transformation is realized in this platform. Since Putnam's (1960) original formulation, it has become the default view among those in the cognitive sciences who take consciousness seriously. This is not to say that there are not dissenters, — for example, a number of researchers have been seeking to ground consciousness in the intricacies of quantum mechanics (see, e.g., Lockwood, 1989, or Stapp, 1995) — but rather that in a still hazy mélange of theoretical conjecture and uncertain empirical footing functionalism seems to most to provide at least a semblance of solid support on which to build a more concrete conception of mind.

There are three primary advantages of functionalism over physicalist accounts. First, by identifying consciousness not with physics but with process, functionalism avoids the chauvinism inherent with physical accounts. If one, for example, claims that consciousness can be identified with a particular substance in the brain, then any entity lacking this substance will also by hypothesis lack consciousness. But this seems to tie consciousness too closely to a particular implementation. Why should we conclude a priori that silicon based

life forms, or even computational machines could not achieve full awareness? Physicalist accounts need not be limited to substances; they can also include electromagnetic fields. As such fields are ubiquitous, the problem of chauvinism is largely relieved. However, this quickly leads to the opposite problem, that of liberalism with respect to sentience. A still glass of water, because of Brownian motion, contains large numbers of fluctuating fields. It is not clear, in this view, why the glass should not be densely populated by the coming and going of thoughts and sensations. What is needed is a theory that avoids panpsychism and at same time excessive chauvinism; functionalism seems to occupy the fertile middle ground (although see Block, 1980, for arguments to the contrary).

The second reason for favoring functionalism over it contenders, and perhaps the driving reason for its popularity is the natural correspondence between conscious and cognitive states.

The latter are usually explicated in functional terms. As a paradigmatic example, consider the processing associated with the Necker cube and similar bistable visual experiences. The working model of this effect posits a competition between the two views. Bolstering this position is the fact that anything done to make one view or the other more salient will give that view a greater chance of winning the initial competition. At the experiential level, something similar is happening; there is a competition with a single winner for the two views. The correspondence between the functional and phenomenal is not exact, because we are generally not aware of the intermediate state between the views. But in this case, and arguably in the vast majority of high-level cognitive processes accompanied by conscious content, an explanation of the general nature of the contents of experience can be had by piggybacking onto the functionalist account, possibly with some additional restrictions.

A final reason for the success of functionalism, and one that will be the most critical for the claims of this paper, is that it singles out causality as the key independent variable for any future full-fledged scientific account of consciousness. The first useful purpose that causality serves is that it helps explain why neural firing is a necessary condition (although not a sufficient one as binocular rivalry and other considerations indicate) for conscious experience. Simply put, a non-firing brain is a dead brain, and more specifically any experiences subserved by a brain module will cease if that module is lesioned. Causality-based theories explain this simply: when firing ceases, causal relations between neurons cease. In contrast, any theory that revolves around a static physical conception cannot handle this

elementary relation between neural activity and consciousness. If a physical account is to be maintained, therefore, a non-static theory must be posited, such as one that invokes energy fields. But this is problematic for a different reason. Suppose we were to press together two active brains. The electromagnetic fields would then overlap, but we would not suppose the conscious contents of the brains would likewise intermingle. In contrast, causality yields a simple explanation of the separation of consciousness. Conscious entities are demarcated by the fact that they are self-contained causally-closed systems. The full implications of this idea are developed further below in the context of the reciprocal corollary.

In summary, there is good reason for believing that if the mind supervenes on the brain, it does so by virtue of the functional properties rather than the physical properties of this organ. More formally, functionalism implies that the constructor C acts on causal variables that transform inputs to behavior to produce qualia that are functions of these variables. In the next section, we consider the central problem with this view from a scientific perspective, namely that no elegant formulation of C exists that is consistent with this notion.

The Inelegance of Functionalism

There is no shortage of conceptual attacks on functionalism. I will briefly mention three forms of attack here:

(a) *Counterintuitive implementations*
Functionalist accounts by construction are independent of physical realization, but this implies, for example, that a collection of doorknobs with two states, open and closed and influencing each other in the appropriate way could have a full array of qualia including pains, emotions, and sensory perceptions (Block, 1980)

(b) *Inverted and absent qualia*
As has been previously argued, there is no necessary connection between qualia and their physical basis. Therefore it is conceivable that when you see red I see green and vice versa. It is also conceivable the color spectrum is inverted, or for that matter, to be a phenomenal zombie, that is, lacking entirely in an inner life but appearing to behave as if one in present (Lycan, 1973; Block, 1980).

(c) *The knowledge argument*
A brilliant neuroscientist is fitted with glasses at birth that

> desaturate all visual input. She learns everything there is to know about vision, but one day takes off the glasses. Arguably, she now knows more about vision when she sees the world in all its Technicolor glory. But how could this be, if color is full explicated by the functionalist account which she already knew (Jackson, 1986)?

These arguments have been thoroughly discussed elsewhere, and I will not do so here, other than to note that the force of these arguments are at least partially blunted by the move to a nomological functionalism; for example, if it is a law of the universe that a given functional transformation produces a red percept, it will not be the case that it could also produce a green or absent quale. Rather, I wish to take seriously that idea that functionalism could in principle provide a scientific account of consciousness, and examine the consequences of that assumption.

One important heuristic that is often overlooked in purely philosophical treatments of functionalism is theoretical parsimony (but see Chalmers, 1996a). It is usually not sufficient that an account be coherent and consistent with intuitions, it must do so by minimizing complexity and maximizing elegance. Conversely, if a particularly simple account is produced that violates intuition, it is sometimes the wise course to table the intuition rather than the account itself. Arguably, this is the path that physicists took with quantum mechanics, which has an elegant formalism but still has yet to fully explicate the conditions under which waves collapse into particles.

In the case of consciousness, there are two motivations for achieving parsimony. First, as in all theoretical endeavors, the simpler theory is more likely to generalize to novel data. A highly disjunctive theory that is 'jury-rigged' to fit a particular set of observations will have difficulties when new empirical results come in. Perhaps a stronger motivation derives from the assumption that consciousness is one of the fundamental quantities in the universe. We are not, for example, trying to derive an elegant description of the liver, which has been provided by evolution with a variety of functions and accompanying mechanisms. Consciousness is most probably something more like gravity, and accordingly, we should be ultimately aiming for something with the elegance of general relativity.

In the terms of the previous discussion, the task is clear: to find a constructor C that is as simple as possible but is still consistent with the data on consciousness. To show that functionalism as traditionally construed cannot achieve this, let us begin by formalizing the notion

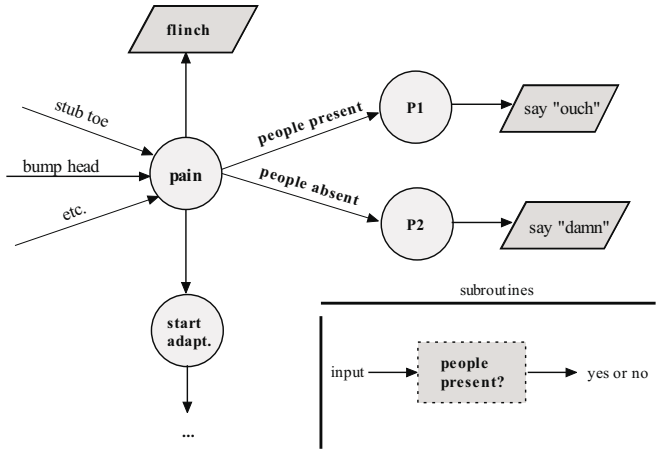

Figure 2. The finite state machine associated with the processing of pain. The appropriate inputs cause a flinch, an adaptive response, and one of two verbal responses.

of a transformation from inputs to outputs. The formalism that will be used is a parallel finite state machine with subroutines (psFSM).[1] A psFSM consists of a set of states, and a set of transitions between states. These choice of which transition to take may be driven either by sensory inputs or by the results of another psFSM that is called as a subroutine. Furthermore, as in the brain, more that one path in the machine may be pursued in parallel.

To make this more concrete, let us examine the hypothetical (and vastly simplified) machine in Figure 2 that describes the algorithm associated with the processing of pain. A number of possible sensory antecedents can trigger the state labeled 'pain'. Then in parallel, three things happen. First, flinching behavior occurs. Second, the pain state causes an adaptive response to initiate, which presumably will reduce the probability of the actions that led to the pain. Third, a subroutine is called (the details of which are not shown) that determines whether other people are in the room. This drives one of two states, which are responsible for the verbal outputs of 'damn' and the more polite 'ouch'.

---

[1] This formalism provides a better intuitive correspondence with the notion of an algorithm than a Turing machine, which is sometimes used to explicate functionalism. The reason is that the Turing machine needs to write any partial results to a tape, and then needs to go to that position on the tape and read the result at the appropriate time. As in a standard computer language, a psFSM can call on a subroutine to generate such a result, without the awkward intermediate reading and writing steps.

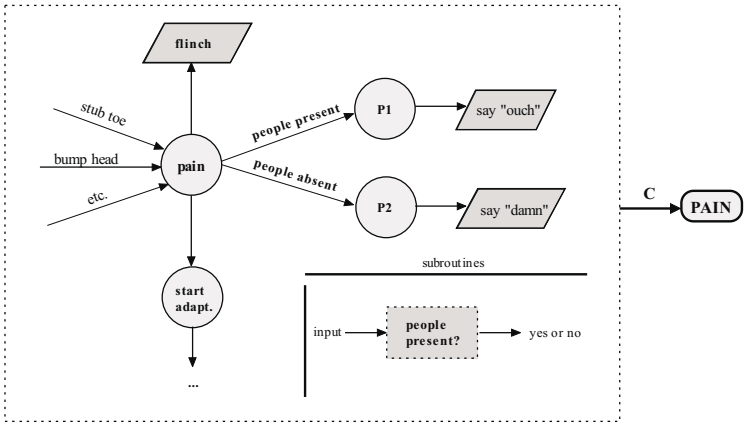

Figure 3. Functionalism states that the psychophysical correspondence transformation C is applied to the process as a whole (inside the dotted box) to produce a given quale.

It is important to realize it is not the states themselves that are responsible for qualia according to functionalism. For example, the 'pain' state is not presumed to be causing the pain; rather it is the causal chain from inputs to outputs and the influences on other internal states (P1, P2, and the adaptation initiation) that is responsible for phenomenal effects. The states themselves have no internal structure, and therefore there is no way that they can be distinguished from one another apart from their relational structure. Another way of saying this is that the label 'pain' has no semantic force; it is provided merely as a means of making sense of the causal chain.

In other words, C must look at the entire sequence of events in order to generate its phenomenal products. This is illustrated in Figure 3. The constructor takes the indicated algorithm, analyses its details, and then returns one or more associated qualia.[2] There are three reasons preventing C from parsimoniously effecting this transformation:

(a) The complexity of the input to C

As a general rule, the complexity of a transformation will be proportional to the complexity of the input to the transformation. There are exceptions to this rule. For example, it may be that a simple and

[2] Do not confuse the two 'pains' in this diagram. The one on the left is a brain state with causal influence on other brain states and behaviors; the one on the right is a quale produced by the psychophysical laws embodied in C. In other words, the one on the right is one that hurts.

well-defined input leads to a lengthy and difficult calculation. Conversely, imagine an enormous database cataloging all Internet purchases of everyone in the world. Some queries to this database will require extensive data mining, but others that concentrate on a few types of purchases (all books bought with the last year, e.g.) may involve a trivially small program. In this case, the search may be confined to a small subset of the input.

In general, however, a complex input to an algorithm implies a certain complexity of processing if only because the input must be effectively parsed before further action is taken. Consider, for example, the transformation effected by C in Figure 3. As previously stated, a central claim of functionalism is that the entire set of causal events from inputs to outputs must be taken into consideration. Therefore C must examine the entire input; by hypothesis, it cannot be pruned to make it simpler. Furthermore, unlike in this simplified example, C must take into account a number of different types of causal consequences of any state. Consider the perception of red. This has consequences for downstream pattern recognition processes, linguistic processes, and emotional processes among other things. The true size of the input to C even in the case of a putatively simple quale would be enormous.

An additional problem with the input to C is that it is semantically charged; i.e., C must understand the meaning of the terms involved in order to make the correct decisions. However, it is notoriously difficult to process semantic information. This is why, for example, search engines still have at best weak natural language understanding. This may be improved in the future, but it is safe to say that there will be never be compact semantic engines. Take the 'people present' decision as a paradigmatic example of the complexity involved. Either C must understand what this means, or it must recursively understand this by reference to the sets of inputs which lead to this decision by the indicated subroutine. The former leads to an intractable or at least very large algorithm, but if the latter tack is taken, then an equally difficult problem arises, treated next.

(b) The inputs and the outputs of the process

One way of viewing functionalism is that it cashes out the semantics of algorithmic processes by tying them down to both the inputs and the outputs to these algorithms. This removes free-floating semantic content within the algorithm although it does produce additional conceptual difficulties. For example, if inputs and outputs are truly important in determining conscious content, then functionalism

inherits many of the problems inherent in behaviorism (Block, 1980). Alternatively, artificially limiting inputs and outputs to ordinary biological mechanisms aligns functionalism at its endpoints with materialism, and thereby inherits the difficulties of this proposal.

Our concern here is as usual the complexity of this aspect of functionalism for the transformation C. Without any boundaries on the nature of the inputs and outputs, it is clear that C faces an intractable task. Simply cataloging the potential set of inputs alone is an unending endeavor. Alternatively, limiting the types of inputs or outputs appears too parochial. Why should a being that sees in the ultraviolet range, that can directly detect magnetic fields, and that communicates via flashing colored lights not be sentient by virtue of these inputs and outputs? But if this is a possibility, a full-fledged theory of consciousness based on functionalism must somehow take these and countless other possibilities into account. In summary, making C a function of world events makes C a function of anything and everything. This is the very opposite of a formula for theoretical elegance.

(c) Counterfactuals in the process

The algorithm illustrated by the flowchart in Figure 3 contains a conditional branch that depends on a binary result generated by a subroutine. However, for any given run of the algorithm, one branch or another will be taken. Maudlin (1989) has argued that this presents serious problems for functionalism in the sense that the contents of consciousness at any given moment will be a partial function of how the brain processed similar stimuli in the past. Bishop's (2002) defense of Putnam's (1988) claim that any physical system, such as a rock, will implement every finite state automaton (FSA), and therefore functionalism (absurdly) implies panpsychism, makes an alternative argument. He considers inter alia holding the input to a FSA constant, such that one path through the state space is always taken. If the FSA generates phenomenal content to begin with, it must also do so after its inputs are constant, as it behaves in an identical fashion, and therefore the counterfactuals (non-taken transitions in the FSA) can't matter with respect to such content.

Once again, the current argument is not concerned with whether counterfactuals could count, but rather with the complexity introduced assuming that they do count (*pace* Putnam/Bishop but consistent with Chalmers, 1996b). It is wildly unparsimonious to assume so, simply because the constructor C must then be a function of past events, and possibly worse yet, of future events in which the

alternative state transitions are taken; otherwise, it has no way of knowing what these alternative transitions are. At a minimum, C must possess an extensive memory to record past events so as to reconstruct a picture of the entire FSA whenever one or another path through the automaton is taken. Moreover, it must remember all combinations and then somehow reconstruct the entire FSA from the bits and pieces it has observed previously. Aside from what appear to be insurmountable conceptual difficulties, such as the fact that current conscious content possibly supervene on future events, this introduces an considerable computational burden on the constructive process.

Taken together, these three difficulties make it highly unlikely that there is an elegant theoretical formulation in which the mind can be conceived as the software of the brain, to invoke a commonly-used description of functionalism. Can it be an accident that functionalist scientific accounts, except at a high level of description (e.g., Baars, 1988), are still thin on the ground? In the next section, an alternative, reduced functionalism is introduced, that attempts to remove these barriers to theoretical success.

Fixing Functionalism: Theory rF

The goal of this section is to introduce a reduced functionalism (rF) by sequentially addressing each of the individual difficulties that arose in the prior section. Later sections will evaluate this pared-down theory for their correspondence with existing conceptions of consciousness and consistency with the empirical data. The following three simplifications correspond respectively to the problems introduced in the previous section:

(a) A bare bones causality

As previously argued, one of the primary advantages of functionalism is that it shifts the explanatory burden from substance to causal relations. One way of doing so is to work with the causal transformations from input to outputs that are effected by an algorithm. However, an algorithm is too unconstrained a representation on which to build a compact theory. The natural alternative then is to retain the essential component of causality and jettison the superfluous aspects of the algorithmic process that serve to complicate the life of the constructor C.

This can be done by looking only at causal relations without reference to a notion of the kinds of processing the brain is performing. For the purposes of this discussion, we will speak of the causal

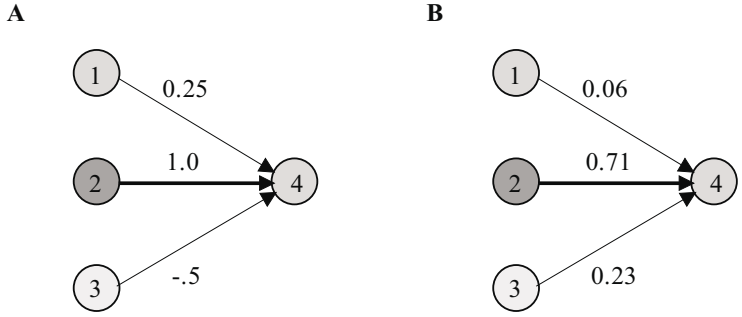

Figure 4. An illustration of the calculation of current flow. A A simple neural network. Weights are as indicated and the firing rate is proportional to the degree of shading. B The resulting causal network.

interactions between neurons, but leave open possibility that variants of the current theory could more profitably look at current flows over larger collections of neurons. A neuron can be characterized by a firing rate and the synaptic efficacy between it and the neurons it influences. Let us designate the causal current between two neurons as the product of this firing rate and the synaptic efficacy, relative to all the other influences the target neuron receives.

Figure 4 illustrates this process in for a simple case. Let us say we wish to know the current from unit 2 to unit 4, and let us assume that relative firing rate of unit 1 is 0.5, that of unit 2 is 0.75, and that of unit 3 is 0.25. Then the relative contribution to unit 4's state by unit 2 is just 0.75*1.0 / (0.25*0.25 + 0.75*1.0 + 0.5*0.5) = .71.[3] The causal network with all influences calculated is shown on the right of Figure 4.

The net result of performing these calculations is a graph, or network, with the nodes of the network representing individual neurons and the edges of the network representing causal influence (not weights as in a typical neural network). This network, while still large, is a considerable simplification over an algorithmic representation. There are no subroutines, and there is no semantics, explicit or implicit, associated with each node.

[3] For the purposes of simplicity, the contribution of the inhibitory weight is assumed to be proportional to its magnitude. Hence the absolute value of this weight is used.

(b) Eliminating inputs and outputs

For the reasons previously stated, there is simply no way of constructing a parsimonious theory of consciousness that depends on the nature of the inputs and outputs to the system. If they cannot be delimited, then the arguments to C are every possible world state. It is safe to say that there are no parsimonious theories with unbounded operands of this sort. Alternatively, limiting the nature of these states seems both arbitrary and parochial.

Thus, if these elements are unwieldy, and there is no way this can be remedied, the only choice is to eliminate them from consideration if parsimony is to be maintained. Ultimately, the survival of the resulting theory will depend on its concordance with the experimental data. However, it is worth noting that this elimination simply means that mental states supervene on brain states and processes, and nothing else[4], which is not an unreasonable *a priori* position.

(c) Delimiting time

The necessity of examining counterfactual causal paths arise if one conceives of consciousness as a function of algorithmic-like processes. Alternatively, one can say that the standard account of functionalism is temporally-independent, and that this permits mental states to be a function of prior mental processes as well as current ones. Apart from the counterintuitive nature of this proposition, this introduces complexities because C is not restricted in its temporal vision. The parsimonious alternative is to have C operate in the present only. Again, this is an assumption that one would probably want to make anyway if one were not under the sway of the computational paradigm. We have no reason to believe that mental states supervene directly on past events, except to the extent that these events modify the brain and therefore affect current brain processes.

Thus, in common with 'standard' physics, rF computes its results as a function of forces at a given instant. In practice, this must be modified slightly when dealing with a neural realization of the theory. At any given time, a neuron is either firing or not, whereas causal

[4] If consciousness is truly a function of causality, as the previous constraint suggests, then it could conceivably be a function of causal events external to the brain as well as those internal to the brain. However, in this case it would not be a function of the way in which the brain transforms inputs to outputs, as computational functionalism suggests; rather external events causally interacting with brain processes would have equal status with regard to phenomenal generation to causal currents inside the brain (cf. the active externalism of Clark and Chalmers, 1998). Furthermore, as will be shown in the next section, in most cases these external causal currents can be eliminated because of the lack of reciprocal interactivity with causal events internal to the skull.

influence in a network is a function of the firing rate. Therefore, in some cases it may be prudent to sacrifice some theoretical elegance and consider a small time window before the present when calculating causal currents.

Summarizing to this point, the constructor C examines the set of quasi-instantaneous causal currents in the brain and from the network of these currents generates the appropriate qualia. Two additional points are in order with respect to the reduction present in rF. First, the fact that C does not operate on counterfactuals, or inputs for that matter, to produce conscious content does not mean that the contents of consciousness are not counterfactually rich, in the sense that this content is sensitive to the input to the system. In fact, given that the argument to C is a causal network, and that the latter can easily engender chaotic dynamics, the contents of consciousness can exhibit extreme sensitivity to small changes in the input space. Second, a similar concern as that raised by Putnam (1988) may arise in this instance. If a rock implements every FSA, by Putnam's state grouping argument, and consciousness is a function of the computation carried out by FSA's, then every object, animate or otherwise, would also contain every qualia. However, it is not the case that a rock implements every causal network. We may find some causal isomorphisms in the rock to simple networks, but if the network is sufficiently complex, then the temporal constraint makes it unlikely that a match would be found — all of the appropriate relations would have to be present at the same time. Thus, by constraining functionalism we are also constraining the set of objects in the universe that are capable of sustaining mental life.

Before proceeding, it is also worth examining the difference between this proposal, and one that it superficially resembles, that Churchland (1986). This work drew on a large number of advances in the neurosciences and the then emerging field of connectionism to identify mind with the workings of appropriate collections of neural networks. There are three primary differences between this conception and the current work: a) the current proposal is still functionalist, although not explicitly computational, whereas the Churchland thesis is closer to that of an identity theory, b) neural networks have been used to explicate the current thesis, but there is nothing explicit that constrains the generators of causal currents to such networks, and c) a key component of the proposal, to be defended in the next section, is that there are restrictions on kinds of causal flows that will be relevant. This restriction puts the current proposal closer in spirit to the reentrant networks of Edelman (2003), as will be discussed further below.

The Reciprocal Corollary

On the basis of what has been said so far, C must take into account all causal currents in the brain in a restricted time window in order to produce a set of corresponding qualia. However, it is not difficult to show that there is a further restriction over and above those suggested by the principle of parsimony which will prove invaluable in helping to verify rF. This is that the only currents that matter are reciprocal, that is, when there is mutual influence between neurons or sets of neurons.

To clarify: let us call the set of neurons that are responsible for the generation of consciousness the conscious kernel. Suppose a neuron A is in that kernel (see Figure 5). Further suppose that another neuron A causally influences B, and no other neurons, but is not influenced by it, either directly or indirectly via intermediate neurons. Then A can have no direct influence on conscious content.[5] Likewise, suppose D is influenced by another neuron C, but D has no influence any neuron in the kernel. Then D can have no influence on conscious content.

Taken together, these two constraints imply the following. There will be a set of neurons reciprocally connected to each other. In the network, this is equivalent to saying that starting with any neuron in this group, one could traverse along the appropriate set of edges and come back to the same neuron. Neurons external to this kernel are purely afferent (inputs) or efferent (outputs) and do not have this property. By the arguments below, such neurons can play no role in the generation of qualia.

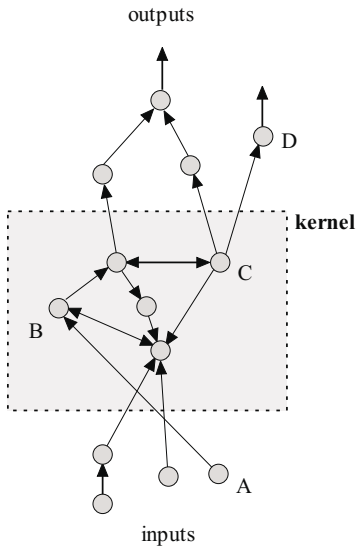

Figure 5. In the kernel, there is a path from every unit to every other unit, including a cycle back to the starting unit itself. Units external to the kernel do not have this property.

[5] By direct influence is meant that the constructor C need not examine the causal currents flowing from this neuron to perform its calculation. This does not preclude indirect influence via other neurons that are in the kernel.

The key to seeing this is the assumption that in any theory that involves causality (both functionalism proper and reduced functionalism) the brain does not occupy a privileged position in the causal chain. Any and all causal influences, both internal and external to the brain are fair game for analytical treatment. In particular, examine the causal path of a light beam from events 1 to 2 to 3 before they hit the eye in Figure 6A. We know that the influence that this beam has on the contents of consciousness is determined solely by its nature when it hits the eye at 3; nothing that happens before this matters. Likewise, we can make similar arguments for the other senses. Whatever is experienced is experienced by virtue of the firing rate of the transducers for these senses, and the causal path taken before they are activated is irrelevant. But, and this is the essential point, if this is true externally to the brain, it is true internally also, assuming that consciousness is determined by causal currents. Therefore feedforward influences by afferent neurons in the brain without feedback to these neurons have no part to play in the generation of consciousness.

Figure 6B illustrates a supplementary argument that demonstrates the same point. If feedforward causal influences had anything to do with consciousness then a light beam that was observed by two brains 1 and 2 would imply that these brains would have overlapping conscious contents, at least to a small degree, because these contents would be a function of this shared path plus whatever happens in the brain. Now as a matter of fact, if these brains are constituted similarly, for example if they belong to two people or even a primate and a human this could be the case. Barring this, however, there is no necessary overlap between the conscious contents of the two observers. For example, if we find life elsewhere in the universe, we would have absolutely no confidence that these beings would have anything like our inner lives simply on the basis of sharing inputs with similar causal histories. Causal antecedents cannot tell us about the inner life of beings; for that we need to look at inner processes.

In the same way, it is easy to show that the causal consequents of any behavior can have no influence on conscious content. Suppose, as in Figure 6c that the pressing of a button leads to one of two quasi-instantaneous actions (unobserved by the presser): event 1, in which a light bulb comes on in the next room, and event 2, in which a clown pops out of a jack-in-the-box in Katmandu. Clearly, the button presser's thoughts will not be influenced by what happens. As before, if this is true outside the brain, it must be true inside also, if only causal currents matter. Thus we can conclude that efferent influences from

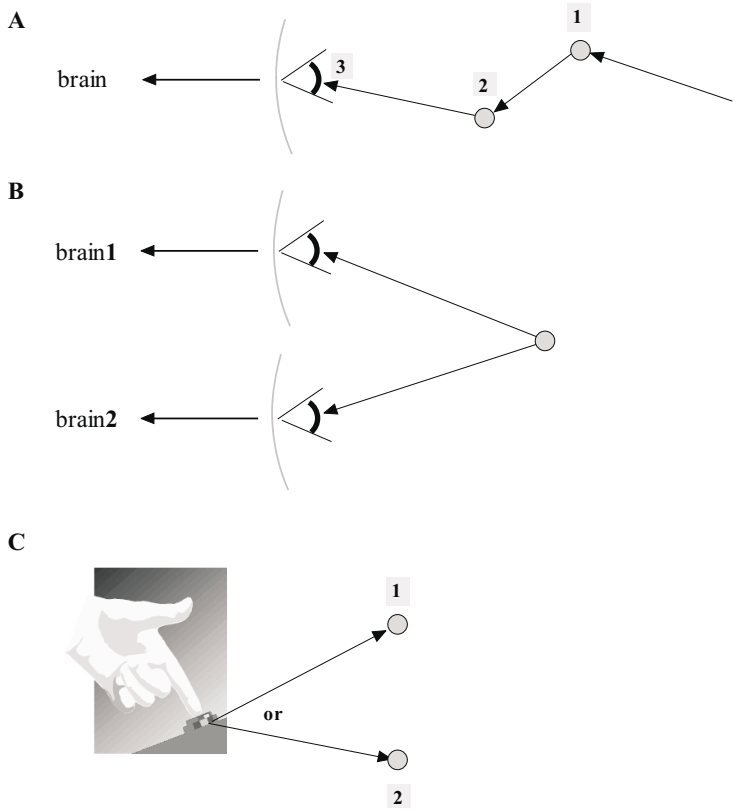

Figure 6. Illustrations of the arguments for the reciprocal corollary. A The causal path of a light beam that enters the eye is irrelevant to the contents of consciousness. B Two brains sharing inputs with similar causal paths have no necessary overlap in their conscious contents. C The causal path taken by the consequents of any behavior is irrelevant to the contents of consciousness.

the neurons that are responsible for consciousness without the additional feedback into this kernel can be ignored.

In summary, if consciousness is a function of causal currents, then all such currents, both inside and outside the brain need to be entertained as arguments to the constructor C. However, we know that the path of the current before it enters the skull is irrelevant to conscious content, and likewise the path after it leaves the body is also irrelevant. Given that causal currents are the only arguments to C, then it

must be the case that all purely afferent and efferent causal flows are irrelevant to the production of consciousness, whether these currents are internal or external to the body. Adding this to the prior theory yields a total of four additional constraints to those traditionally contained in the functionalist thesis:

(i) *The representational constraint:*
Only the network formed by looking at causal currents matters;

(ii) *The I/O constraint:*
The inputs to the network and the outputs from this network can be ignored;

(iii) *The temporal constraint:*
Only currents with a narrow time window centering on the present are considered; and

(iv) *The reciprocal constraint:*
Vertices in the network that are not reciprocally connected to other vertices in the network (in graphical terms, are not part of a cycle) are pruned from consideration.

In what follows, I will attempt to show that this reduced functionalism[6] is similar in many respects to other explanatory frameworks that have been advanced and is sufficient to explain the broad character of some experimental results in this area. But before doing so, let us examine the putative simplicity of rF relative to computational functionalism in more detail.

Compositional phenomenology

Theory rF postulates that the input to the constructor is a graph of causal currents rather than an algorithm. While the former is conceptually simpler than the latter by a wide margin, as it contains no semantics, either intrinsic or in relation to the inputs that drive it, the sheer size of the brain endows it with a numerical complexity which if not addressed would obviously be fatal to theory that aims for parsimony. The purpose of this section is to show that this complexity can be managed by breaking the graph down into constituent components.

---

[6] Whether such a theory still falls under the functionalist rubric may be questioned. However, rF like functionalism proper is still a theory about dynamics rather than substance, although it concentrates on immediate causal relations rather than the process these relations may be realizing. For this reasons, perhaps the clearest label is that of a non-computational functionalism.

First, a relatively minor point: the restriction of the graph vertices to those in the kernel, that is, those cells that are reciprocally interacting in a given time window will significantly reduce the size of the graph. The non-reduced graph has approximately 1011 vertices, and 103 edges per vertex corresponding respectively to the number of neurons and the number of synapses per neuron for a total of 1014 edges. Let us make a generous assumption and claim that only 1 in a 100 cells will be in the kernel at any one time. Much of this reduction will take place due to fact that sparse encoding in early sensory processing is responsive to only a single feature in a single region. For example, most of the orientation detectors in primary visual cortex will not be active in response to a given visual input.

The problem is that this leaves us with a graph with 109 vertices and on the order of 1012 edges. If the constructor were forced to observe the graph en masse in order to produce the proper phenomenology, it would still face a task of enormous complexity. However, this ignores the possibility that the graph may be naturally decomposable. The constructor can break down the graph into constituent parts, process these separately, and then stitch the micro-qualia, as it were into the resulting phenomenal field.

As an example of this process, consider the construction of the visual perceptual field. Processing in primary visual cortex V1 proceeds by repetition of regular structures, with little variation between these structures, and in such a way as to preserve the topography of the input field. This is also true for downstream areas V2-V5, although these areas will have larger receptive fields than V1, and also are specialized for feature content; for example, V4 is thought to underlie the processing of color, V5 (MT) the processing of motion. We can conceive of a virtual column or partition to process a phenomenal 'pixel' running from V1 and including the corresponding cells in higher processing areas (and possibly area IT) that send descending feedback to the corresponding earlier modules. We can also conceive of a horizontal set of connections between such partitions that are responsible for different processing effects depending upon the processing module, for example inhibitory lateral connections subserving color contrast and color constancy in V4 (1996) and contour integration in V1 (Li, 1998), and in this context, act as the glue that holds the elements of the visual perceptual together. Thus cells dedicated to a pixel would be far fewer than that for visual processing as a whole, roughly by six orders of magnitude assuming on the order of 106 pixels in the foveal area of

attention. The resultant graph would then have 109/106 = 103 vertices, or possibly less by pruning regions of insignificant causal flow.[7]

In summary, although the constraints imposed by rF on the input to the constructor will still produce a large and unwieldy network, it may be possible to decompose this network into regular repeating parts, and construct the phenomenal field from these. I am not claiming that any aspect of understanding this process is trivial. However, the theory required to explain qualia as the result of a causal network is, if you like, comparable in complexity to that of explaining behavior as the result of a neural network. That is to say that it is a difficult endeavor, but well within the purview of the scientific method. Before leaving this topic, it is worthwhile mentioning a few requirements for a full-fledged account along these lines:

(i) *Graph partitioning*
The constructor C must follow regular rules when deciding how to partition the graph into constituent parts (unlike us, it does not have a priori knowledge of the structure of the visual system). One possibility is that within a partition, cells are strongly interconnected, and between, less so. The picture is similar to the kind of network that is produced when mapping paper citations (or Facebook friends for that matter); academic fields and subfields look like tight balls, with thin lines connecting these disciplines.

(ii) *The qualia of a partition*
Once a partition is isolated, the task will be to correlate the graph within that partition with the corresponding phenomenal content. Confidence in this endeavor, as in 'normal' science, will be engendered when regularity results, that is, when distances between graphs as given by a suitable metric entail differences of equal or approximate magnitude in phenomenal space.

(iii) *Other modalities and aspect of consciousness*
The spatial nature of vision makes it naturally partitionable; it remains to be seen if a similar approach can be used with other sensory modalities, and with emotional or other less well-defined feels.

[7] Here we keep the 109 vertex estimate even though we are only looking at visual processing because of the large proportion of the brain dedicated to vision.

Theoretical Correspondences

The ideas presented here are not entirely new, although the specific route taken in generating them via the notions of parsimony and the reciprocity corollary may be novel. There are at least three related theories that either correspond to rF or to which rF reduces as a special case. The first such theory has a long history (see Pollen 1999 for a review), and specifically identifies resonant loops between processing modules in the brain as a necessary condition for consciousness (cf., Grossberg, 2001). Edelman, in particular, has been one the most forceful advocates of the view identifying reentrant loops with phenomenal activity (2003). By reentry he means that higher-order categorical knowledge in the frontal, temporal and parietal areas interact with primary and secondary cortical areas in a feedback loop. There are four claims associated with this view: (a) that such bidirectional feedback can provide superior pattern recognition and other forms of processing, (b) that reentry facilitates feature binding, (c) that reentry allows for information to be broadcast across modules, and (d) that reentry is necessary for consciousness.

The first claim is computational as opposed to phenomenal and will be ignored here, and the middle two claims will be considered separately below; here we concentrate on the correspondence between rF and reentry with respect to consciousness. It is easily seen that reentry and reciprocity are closely related concepts, the key difference being that the former involves interactive feedback between modules rather individual neurons. However, recall that rF is neutral with respect to granularity; causal flow may be described between neurons or sets of interacting neurons. Furthermore, Edelman's concept of a dynamic core, or the subset of neural activity involved in reentrance, and alone responsible for the generation of consciousness, is closely related to the notion of the conscious kernel introduced above. In summary, although different justificatory routes have been taken in their production (the current work concentrates more on first principles), in both rF and Edelman's account consciousness is not the result per se of computations in the brain, but rather the set of immediate and interactive causal currents.

An associated claim is that reentry facilitates temporal binding via mutual excitatory feedback. In standard neural models of pattern recognition, the disparate elements of a stimulus are assumed to become bound together through convergence, such that in each successive level of processing there are recognition units for groupings of features from the layer below. However, this leads to the well-known

problem of combinatorial explosion (von der Malsburg, 1981), because it implies that there would need to be higher-level units for each combination of lower-level features. An alternative solution is temporal binding (Engel *et al*., 1999), whereby synchronized firing between neurons subserving the elements of the stimulus is hypothesized to form the basis for the grouping of features. This suggestion has received evidentiary support single cell studies in the cat and EEG and MEG studies in humans (see Engel and Singer, 2001 for a review). Crick and Koch (1990) have also suggested that synchronized firing cycles in the 40 Hz range may act as both as a computational mechanism to avoid the problems of convergence and means of ensuring binding in awareness.

It is the latter claim that we wish to examine here. Recall that one of the constraints introduced both to eliminate counterfactuals and to give rF the flavor of a standard scientific account was to make consciousness a function of contemporaneous causal forces. rF provides a strong hint as to why firing synchrony implies phenomenal binding; it is simply because the mental supervenes on the physical plus the temporal. In other words, from the point of view of the constructor C, two neurons sequentially influencing each other in successive time windows is equivalent to there being no reciprocal influence at all; only reciprocal currents in the same time window count. If this window is on the order of 25ms (or smaller), then rhythms in the 40Hz range can be accounted for. In summary, if rF were correct, one would expect that firing synchrony would be a paramount factor in determining which stimulus elements are joined together phenomenally, and which are entertained in separate phases of awareness.

A final correspondence may be drawn between rF and theories suggesting that, using Dennett's (1993) terminology, the contents of consciousness enjoy a degree of cerebral celebrity. That is, once an item enters in the conscious arena, regardless of its origin, it becomes accessible to all other processing modules. A closely related notion is Baars' (1988) cognitive theory of consciousness, in which working memory plays the role of central store, and the contents of consciousness are identified with its contents. Once in working memory, an item may be then modified by other modules operating beneath the level of consciousness, and then optionally placed back in the central store. Both models gain impetus from long-standing cognitive models such as production systems (Klahr *et al*., 1987), as well as the informal introspective observation that once an item is conscious, it seems that we can act on it at will — it can form the basis of a behaviour, such as a speech act, or it be drawn upon to make further inferences.

Theory rF and in particular the reciprocal corollary also predict that accessibility will be a property of consciousness. Consider that to be in the conscious kernel implies that, at any given instant, there is a route from every element in the kernel to every other unit. Furthermore, if one assumes a model such as that of Baars, in which processing modules revolve around a central store, this route will be direct in the sense that each module will place its contents in this store, which will then be one step away from every other module.

An outstanding question regardless of implementation is the whether all conscious contents are 'broadcast' to the rest of the brain, and if so, this is a contingent fact or a necessary property in order for the item to become conscious in the first place. Here too rF may be able to provide some insights. One counterexample to complete 'celebrity' would be when two or more distinct kernels arise in the same brain. Presumably only one would have access to the linguistic center of the brain, and only one would be reportable, but there would still be two non-communicating consciousnesses. Tononi and Edelman (2001) make a similar point with respect to their model of schizophrenia, and others (e.g., Sperry, 2001) have made similar claims with respect to split-brain patients. Severing the corpus callosum immediately removes the possibility of any causal contact between the hemispheres (modulo remaining fibers), and rF would therefore predict the development of separate consciousnesses.

Empirical Correspondences

Here we consider three strands of evidence consistent with the fact that consciousness depends on quasi-instantaneous recurrent causal interaction. Of necessity, complex phenomena with often conflicting views and large sets of results will be succinctly presented, but the hope is that enough detail will be given to make rF at least a plausible foundation on which a more detailed theory can later be constructed. The first such strand derives from the well-known practice of suppressing the appearance of a visual stimulus from awareness by presenting another masking stimulus within 50 to 150ms of the original stimulus. This paradigm is used for example to test the role of priming without the mediation of awareness.

Figure 7A shows the standard backward masking experiment. Here a target is followed by a center-aligned mask in rapid succession. Depending on a number of variables including luminance, distance between target and mask, stimulus onset asynchrony and others, the mask will suppress the appearance of the target. One feedforward only

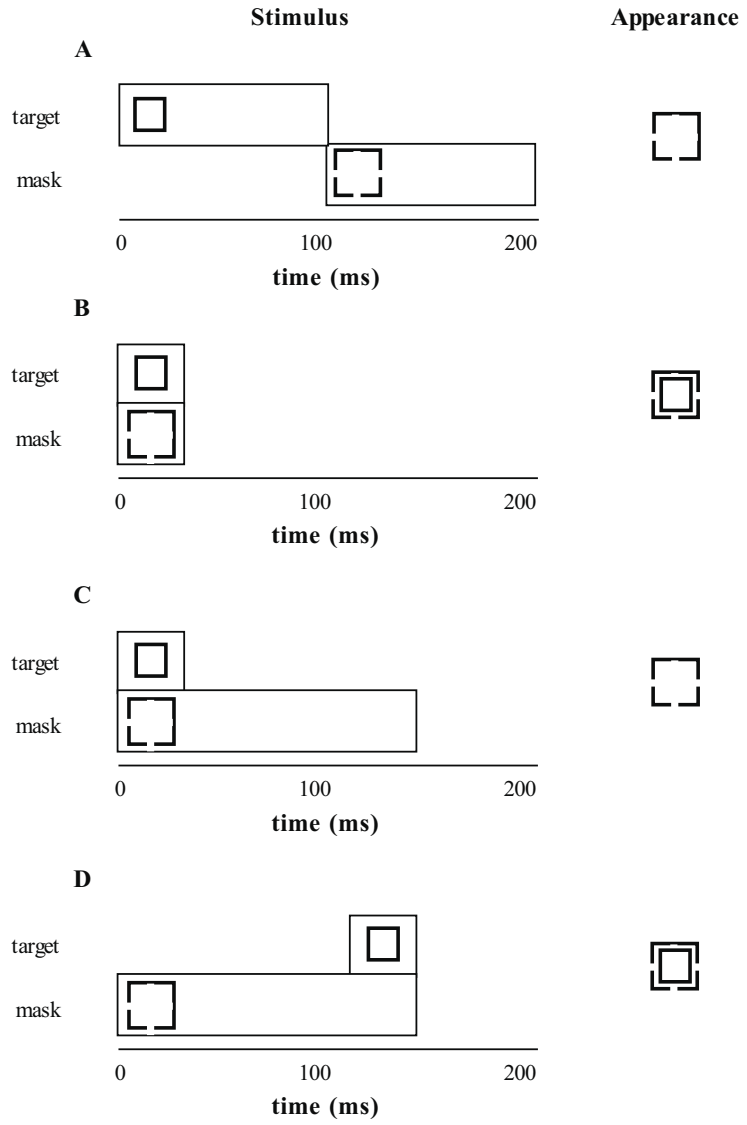

Figure 7. Variants on the masking paradigm. A The standard backward masking experiment. B The target and mask have the same offset and onset. C The mask has a later offset than the target. D The target and masks have the same offset but the mask has a prior onset.

explanation for this effect is that the onset of the mask inhibits the reverberation of the target in memory in higher (extrastriate) areas of visual processing in the brain (Kolers, 1968; Turvey, 1973). However, this account predicts that optimal suppression of the target is an inverted U-shaped function of mask-target stimulus onset asynchrony (SOA). When SOA is 0, as in Figures 7B and 7C, it gives equal weighting to the stimuli and both should be seen. This is indeed the case in 7B, but not so in 7C, in which the mask is allowed to remain after the removal of the target (Di Lollo, Enns, and Rensink, 2000). In this instance, this perseverance wipes the target from perception. One way of rescuing the inhibitory account is to claim that the competition between stimuli will be a function of presentation time. However, this fails to explain the results shown in 7D, in which the longer presentation of the mask does not affect the target, as long as they have the same or nearby offsets.

Di Lollo, Enns, and Resink (2000) provide an alternative model in which conscious visual activity is the result of a match between bottom-up perceptual signals in the primary visual cortex and top-down contextual signals from secondary visual modules. This model explains standard backward masking as the mismatch between target signal which originally triggers higher-order object recognition signals in secondary cortex and the incoming mask. In the case of Figure 7B, both signal and target reach higher processing in synchrony, so no mismatch occurs (the existence of iconic memory is assumed in primary visual cortex to allow the signals there to persevere and match the top-down signals, which typically will not be generated until after 80 ms). In Figure 7C both the target and mask move on to secondary processing, but only the mask remains in primary memory, and thus only it is perceived (there will be no match for the target). In 7D, the mask reaches secondary processing first, but the target then comes along to join it. Both are present in primary memory with equal intensity, thus they match the corresponding top-down signal, and both are perceived.

These considerations suggest that neither primary nor secondary visual processing alone is sufficient for visual consciousness, but rather it is the real-time interaction between the two that leads to perception, in accord with the theory rF. It is difficult to explain the collection of results in Figure 7 by feedforward processing alone (although see Francis and Hermens, 2002). Rather, it is the resonance between bottom-up and top-down processing that allows a stimulus to enter awareness. Lamme (2001) also argues that the temporal characteristics of masking are consistent with this notion. It takes

approximately 100ms for an image to reach awareness, and this is close to the time it takes for neural activity to spread from primary to secondary visual areas and back again.

A closely related notion to masking is that of contextual modulation. In general, we are not aware of the raw stimulus, but rather the result after a context-driven feedback process. If feedforward activation were sufficient to produce awareness, then one would expect perception to proceed in two stages, first a relatively unprocessed stimulus, and then the stimulus altered by further processing. In fact, we are typically aware only of the latter, and if the stimulus is too brief to engender resonance between bottom-up and top-down processes, then nothing is seen at all, as in a typical backward masking experiment. In the visual domain, for example, the perception of color is modulated by both the global context, resulting in the phenomenon of color constancy, in which the illuminant is partially discounted (Land & McCann, 1971), and local context, resulting in the phenomenon of color induction, in which neighboring opponent hues serve to increase color saturation (Jameson & Hurvich, 1959). In the auditory domain, top-down feedback colours both normal auditory (Connine, 1987; Ganong, 1980; Mann, 1986) and musical perception (Katz, 2004).

An alternative explanation to resonance for effects such as these is to stipulate that consciousness only has access to the final layers of perceptual processing; this could explain why only the processed stimulus is perceived. One difficulty with this account is that secondary layers also participate in processing, again leading to a potential two stage perceptual process, contrary to perceptual reality. For example, in the both the Wray and Edelman (1996) and Courtney et al. (1995) models of color constancy, lateral inhibition within V4 plays a key role in this effect. In addition, higher-order processing layers are generally at a lower resolution than perceptual acuity would suggest, indicating that perception is generated at least partly from primary processing. Furthermore, there is direct evidence that primary cortical activity participates in the construction of conscious content. For example, Lamme, Zipser, and Sprekreijse (1998) recorded the responses in V1 in awake monkeys to figures on a background. Before 80 ms, neurons responded only to the local quality of the texture in the image. Between 80 and 120 ms, they also responded to figure-ground boundaries. After 120 ms, they responded to the surface of the figure only. After lesioning the extrastriate cortex, V1 neurons reverted to judging local features only, regardless of the length of presentation. A similar result was seen under anesthesia. Taken together, these results

suggest that higher-order feedback modulates activity in primary visual cortex, and that it is this latter activity that reaches awareness.

This paper began with an example of binocular rivalry, in which perception alternates between the two competing images. Binocular rivalry represents one of the best tests of any proposed theory of consciousness because there is a clear and well-defined perceptual change in the absence of stimulus alteration; therefore, if a brain-based process correlates with the change in conscious content it may be a viable candidate for the neural substrate of consciousness. Here three sets of results that bear directly on this issue are summarized:

(i) Initial single-cell recording studies in monkeys (Logothetis & Schall, 1989; Leopold & Logothetic) found that a greater proportion of neurons in later visual areas correlated with the changes in perceptual dominance. However, since this time, a number of considerations have led to the view that the involvement of primary visual cortex may have been underestimated (Polonsky, et. al., 2000). In addition, in an fMRI study in which the images were tagged by using different levels of contrast, Polansky et. al. (2000) found that the correlation between change in V1 and ocular dominance were comparable to those in V2, V3, and V4. Thus, it would appear that all levels of visual processing subserve the maintenance of the dominant, consciously perceived image.

(ii) The next question revolves around whether the rivalry is between the eyes or the pattern. That is, is this a competition in the brain between the differing patterns presented to the two eyes, or does it involve a competition between the separate representations of monocular information? If the former, it would likely implicate secondary areas of visual processing responsible for pattern classification, in support of the original single-cell recording studies mentioned in i); if the latter, earlier areas representing purely monocular information and possibly secondary areas would be implicated. Blake (2001) has argued that except under special circumstances, the rivalry is eye-based. In particular, he showed that if patterns shown in the dominant and suppressed eyes are reversed, the pattern that is now being shown to the dominant eye will be seen. If the rivalry was pattern-based, one would expect no change in the perceived image.

(iii) Most interestingly for the purposes of the current discussion, Fries et. al. (1997) found that the degree of synchronization between neurons in areas 17 and 18 of the cat increased for the selected eye, and decreased for the suppressed eye. They also found no significant differences in firing rates between these two case, suggesting that temporal synchrony rather than activity is critical for visual experience, at least in early visual processing.

While it is still too early to give a definitive account of the neural basis of binocular rivalry, it is worthwhile to note that rF is consistent with these sets of results. First, as suggested by the results in (i) and (ii), it appears that activity in both primary and secondary visual cortices are implicated in the generation of perceived image. This is consistent with the notion that it is the interaction between these and possibly other brain areas that subserve visual experience, rather than any area operating alone. The result in (iii) is consistent with the idea that conscious experience is a function of quasi-instantaneous neural activity. It is unknown whether the synchrony in this case is facilitated by lateral connections within primary visual cortex or via descending influences (or both). In either case, however, it is the in-phase relationship between firing patterns rather than the overall strength of activity that is determining conscious content.

Discussion

The central contention of this paper is that parsimony can be a valuable heuristic in guiding the search for a scientific theory of consciousness. As has been argued, this heuristic transfers the burden of the generation of consciousness from computations to quasi-instantaneous causal currents. Other considerations further restrict these currents to reciprocal influences alone. The resulting theory has been shown to be consistent with a number of existing accounts, and also to explain the broad character of a number of experimental results.

Whether parsimony can be of significant aid in producing a full-fledged theory of consciousness, and whether the specific suggestions introduced in this paper will be helpful in doing so is still an open question. However, to put the current collection of arguments into perspective it will prove instructive to return to the distinction made at the start of the paper between logical and (mere) nomological supervenience. The former implies that there will some chain of reasoning from the physical to the mental, such that the mental could not be otherwise once the physical is truly understood. However, if

Chalmers (1996a) is correct, and the mental does not logically supervene on the physical, such a chain does not exist. Another way of saying the same things is as follows: Science will never have a Eureka moment, in which some clever theorist says 'Aha, now I see why qualia must arise when the brain is doing such and such.' This is a form of explanation-based or deductive learning (Michell et. al., 1986). Explanation-based learning takes a set of facts and an explanandum, shows that a subset of those facts plus other background assumptions entails the explandandum, and concludes that the subset is the cause of the explanandum. But explanation-based learning is possible only if an entailment exists, and by hypothesis, the mental is not derivable from the physical.

Where does that leave the scientific pursuit of a theory of consciousness? Explanation-based learning is the most powerful weapon in the scientific arsenal, but it is not the only one. There is also what may be termed brute-force induction, or inductive learning. To take a simple case, if $x + a \Rightarrow F$, $x + b \Rightarrow F$ and $x + c \Rightarrow F$, a reasonable guess is that x is the cause of F. For example, suppose it was found that the one invariant in visual consciousness was a synchronized interaction in the gamma range (~40Hz) between the frontal cortex and the secondary visual cortices. Given this strong correlation, we may be tempted to make the leap from the mere correlation between gamma activity and consciousness and raise it to the level of a full-fledged causal relation.

Without an explanatory chain, however, saying just why synchronized activity at a certain frequency must correlate with consciousness, this will remain a weak conclusion. For example, if we were to encounter intelligent Martians without such activity, it would be rash to claim that they are phenomenal zombies just because they lack this feature. Likewise, the absence of synchronized activity at this frequency in a machine should not thereby preclude it from sentience, nor will the addition of such activity guarantee sentience.

Brute-force induction is just too weak in general to generate true causal explanations, and it is especially weak in the case of consciousness, because we have too few examples from which to generalize. Therefore, if it is to work at all, it must be augmented. What this paper has suggested is that the notion of parsimony can elevate the weak claims of functionalism to the status of proto-theory or better by trimming the fat from an otherwise untenable account. In the current case, this excess included the unworkable notions inherent in the notion of the algorithm itself, the superfluous role played by the inputs and outputs, and the unbounded time frame in which all this supposedly takes

place. The remaining residue, a graph of causal relations, at the very least has a chance of being proved wrong, because it provides a compact and workable theory.

In summary, the battle lines between science and mystery with respect to consciousness line up as follows. On the side of mystery is the impossibility of explanation-based reasoning, and in addition the not inconsiderable problem of other minds. We have no 'cerebrometer' that tells us what anyone other than ourselves are feeling at any given time, nor could one be built without a firm scientific theory already in place. On the side of science is the possibility of inductive reasoning, but in addition theoretical parsimony to bolster this weak method. It remains unclear which side will prevail in this struggle, although as this paper has argued science stands a better chance once it actively includes the latter heuristic.

References

Baars, B.J. (1988), *A Cognitive Theory of Consciousness* (Cambridge University Press).

Bishop, M. (2002), 'Counterfactuals cannot count: A rejoinder to David Chalmers', *Consciousness and Cognition*, **11**, pp. 642–52.

Blake, R. (2001), 'A primer on binocular rivalry, including current controversies', *Brain and Mind*, **2**, pp. 5–38.

Block, N. (1980), 'Troubles with functionalism', in N. Block, ed., *Readings in the Philosophy of Psychology*, Vol 1. (Harvard University Press).

Chalmers, D.J. (1996a), *The Conscious Mind: In Search of a Fundamental Theory* (Oxford University Press).

Chalmers, D.J. (1996b), 'Does a rock implement every finite-state automaton?', *Synthese*, **108**, pp. 309–33.

Churchland, P.S. (1986), *Neurophilosophy: Toward A Unified Science of the Mind-Brain* (Cambridge, MA: MIT Press).

Connine, C.M. (1987), 'Constraints on the interactive processes in auditory word recognition: The role of sentence context', *Journal of Memory and Language*, **26**, pp. 527–38.

Courtney S.M., Finkel L.H. & Buchsbaum G. (1995), 'Network simulations of retinal and cortical contributions to color constancy', *Vision Research*, **35**, pp. 413–34.

Crick, F. & Koch, C. (1990), 'Toward a neurobiological theory of consciousness', *Seminars in the Neurosciences*, **2**, pp. 263–75.

Dennett, D.C. (1993), 'The message is: There is no medium', *Philosophy and Phenomenological Research*, **53**, pp. 919–31.

Di Lollo, V, Enns, J.T. & Rensink, R.A. (2000), 'Competition for consciousness among visual events: The psycholphysics of reentrant visual processes', *Journal of Experimental Psychology: General*, **129**, pp. 481–507.

Edelman, G.M. (2003), 'Naturalizing consciousness: A theoretical framework', *PNAS*, **100**, pp. 5520–24.

Engel, A.K. , Fries, P. , Konig, P. , Brecht, M. & Singer, W. (1999), 'Temporal binding, binocular rivalry, and consciousness', *Consciousness and Cognition,* **8**, pp. 128–51.

Engel, A.K., & Singer, W. (2001), 'Temporal binding and the neural correlates of awareness', *Trends in Cognitive Sciences*, **5**, pp. 16–25.
Fries P., Roelfsema P.R., Engel A.K., Konig P., & Singer W. (1997), 'Synchronization of oscillatory responses in visual cortex correlates with perception in interocular rivalry', *Proceedings of the National Academy of Sciences*, **94**, pp. 12699–704.
Ganong, W.F. (1980), 'Phonetic categorization in auditory word processing', *Journal of Experimental Psychology Human Perceptual Performance*, **6**, pp. 110–25.
Grossberg, S. (2001), 'Brain learning, attention, and consciousness', in B. Baars, W. Banks & J. Newman, eds., *Essential Sources in the Scientific Study of Consciousness* (Cambridge, MA: MIT Press).
Jackson, F. (1986), 'What Mary didn't know', *Journal of Philosophy*, **83**, pp. 291–5.
Jameson, D. & Hurvich, L.M. (1959), 'Perceived color and it dependence on focal, surrounding and preceding stimulus variables', *Journal of the Optical Society of America*, **49**, pp. 890–8.
Katz, B. (2004), 'A measure of musical preference', *Journal of Consciousness Studie*s, **11** (3), pp. 28–57.
Klahr, D., Langley, P. and Neches, R. (1987), *Production System Models of Learning and Development* (Cambridge, MA: The MIT Press).
Kolers, P.A. (1968), 'Some psychological aspects of pattern recognition', in P.A. Kolers & M. Eden, eds., *Recognizing Patterns* (Cambridge, MA: MIT Press).
Kripke, S.A. (1972), *Naming and Necessity* (Harvard University Press).
Lamme, V.A.F. (2001), 'Blindsight: The role of feedforward and feedback corticocortical connections', *Acta Psychologica*, **107**, pp. 209–28.
Lamme, V. A.F., Zipser, K. & Spekreijse, H. (1998), 'Figure-ground activity in primary visual cortex is suppressed by anaesthesia', *Proceeding of the National Academy of Sciences, USA*, **95**, pp. 3263–8.
Land, E.H. & McCann, J.J. (1971), 'Lightness and retinex theory', *Journal of the Optical Society of America*, **30**, pp. 3–32.
Leopold, D.A. & Logothetis, N.K. (1996), 'Activity changes in early visual cortex reflect monkeys' percepts during binocular rivalry', *Nature*, **379**, pp. 549–53.
Li, Z. (1998), 'A Neural model of contour integration in the primary visual cortex', *Neural Computation*, **10**, pp. 903–40.
Lockwood, M. (1989), *Mind, Brain, and the Quantum* (Oxford University Press).
Logothetis, N.K.& Schall, J.D. (1989), 'Neuronal correlates of subjective visual perception', *Science*, **245**, pp. 761–3.
Lycan, W.G. (1973), 'Inverted spectrum', *Ratio*, **15**, pp. 315–19.
Mann, V. (1986), 'Distinguishing universal and language-dependent levels of speech perception: Evidence from Japanese listeners' perception of English "l" and "r"', *Cognition*, **24**, pp. 169–96.
Maudlin, T. (1989), 'Computation and consciousness', *Journal of Philosophy*, **86**, pp. 407–32.
Mitchell, T.M., Keller, R.M. & Kedar-Cabelli, S.T. (1986), 'Explanation-based generalization: A unifying view', *Machine Learning*, **1**, pp. 47–80.
Polansky, A., Blake, R., Braun, J. & Heeger, D. (2000), 'Neuronal activity in human primary visual cortex. correlates with perception during binocular rivalry', *Nature Neuroscience*, **3**, pp. 1153–9.
Pollen, D.A. (1999), 'On the neural correlates of visual perception', *Cerebral Cortex*, **9**, pp. 4–19.
Putnam, H. (1960), 'Minds and machines', in *Dimensions of Mind: A Symposium*, ed. S. Hook (New York University Press).

Putnam, H. (1988), *Representation and Reality* (Cambridge, MA: Bradford Books).
Sperry, R.W. (2001), 'Hemisphere deconnection and unity in conscious awareness', in B. Baars, W. Banks & J. Newman, eds., *Essential Sources in the Scientific Study of Consciousness* (Cambridge, MA: MIT Press).
Stapp, H.P. (1995), 'Why classical mechanics cannot accommodate consciousness but quantum mechanics can', *Psyche*, **2** (5).
Tononi G. & Edelman G.M. (2000), 'Schizophrenia and the mechanisms of conscious integration', *Brain Research Reviews*, **31**, pp.391–400.
Turvey, M.T. (1973), 'On peripheral and central processes in vision: Inferences from an information-processing analysis of masking with patterned stimuli', *Psychological Review*, **81**, pp. 1–52.
von der Malsburg, C. (1981), 'The correlation theory of brain function', in E. Domany, J.L. van Hemmen and K. Schulten, eds., *Models of Neural Networks II* (Springer).
Wray, J. & Edelman, G.M. (1996), 'A model of color vision based on cortical reentry', *Cerebral Cortex*, **6**, pp. 701–16.

Paper received March 2007