Putting Content into a Vehicle Theory of Consciousness

Gerard O'Brien and Jon Opie

Department of Philosophy
University of Adelaide
South Australia 5005

gerard.obrien@adelaide.edu.au
http://arts.adelaide.edu.au/Philosophy/gobrien.htm

jon.opie@adelaide.edu.au
http://arts.adelaide.edu.au/Philosophy/jopie.htm

Appeared in *Behavioral and Brain Sciences* **22**:175-96 (1999)

Abstract

The connectionist vehicle theory of phenomenal experience in the target article identifies consciousness with the brain's explicit representation of information in the form of stable patterns of neural activity. Commentators raise concerns about both the conceptual and empirical adequacy of this proposal. On the former front they worry about our reliance on vehicles, on representation, on stable patterns of activity, and on our identity claim. On the latter front their concerns range from the general plausibility of a vehicle theory to our specific attempts to deal with the dissociation studies. We address these concerns, and then finish by considering whether the vehicle theory we have defended has a coherent story to tell about the active, unified subject to whom conscious experiences belong.

Introduction

Our target article sets out to defend a way of thinking about consciousness which, while not completely novel, is certainly unfashionable in contemporary cognitive science. Most theories in this discipline seek to explain conscious experience in terms of special computational processes which privilege certain of the brain's representational vehicles over others. In contrast, we conjecture that phenomenal experience is to be explained in terms of the intrinsic nature of the explicit representational vehicles the brain deploys—in terms of what these vehicles *are* rather than what they *do*. Given that vehicle theories of consciousness are rare in cognitive science, we expected our target article to receive a good deal of criticism, and our expectations were not dashed. What is gratifying, however, is the constructive spirit in which this criticism is proffered. If by the end of this reply our connectionist vehicle theory of consciousness is any more intelligible (and dare we hope, any more plausible), it is the commentators we have to thank.

The commentaries raise difficulties and objections across many fronts, ranging from the neuroscientific to the philosophical. It has been a daunting task to structure a reply that responds to all of these worries in a perspicuous fashion. In the end, we have settled on seven inter-related themes, embracing both the conceptual foundations on which our vehicle theory is based (Part I), and its general empirical plausibility (Part II) (see Table 1).

==========

Table 1. Outline of Response

Part I: Conceptual Foundations

1. Vehicles (**Church, Cleeremans & Jimenez, Dennett & Westbury, Kurthen, McDermott, Wolters & Phaf, Thomas & Atkinson, Van Gulick, O'Rourke**)

2. Representation (**Perner & Dienes, Ellis, O'Rourke, Reeke, Mac Aogáin, Wolters & Phaf, Church, Clapin, Dennett & Westbury, Schröder, Lloyd**)

3. Stability (**Lloyd, Mangan, Reeke, Cleeremans & Jimenez, Dennett & Westbury, Gilman, Pólya & Tarnay, Schröder, McDermott, Perner & Dienes**)

4. Identity (**Ellis, van Heuveln & Dietrich, Kurthen, Newton, Velmans**)

Part II: Empirical Plausibility

5. General Empirical Concerns (**O'Rourke, Velmans, Cleeremans & Jiménez, Gilman, Mangan, Perner & Dienes, Van Gulick, Schröder, Mortensen**)

6. The Dissociation Studies Revisited (**Perner & Dienes, Velmans, Dulany, Perruchet & Vinter, Zorzi & Umilta, Kentridge**)

7. Subject Unity, Agency and Introspection (**Carlson, Dulany, Schwitzgebel, Mac Aogáin, McDermott, Perner & Dienes, Coltheart**)

==========

Part I: Conceptual Foundations

Any adequate scientific theory must satisfy multiple constraints of both a conceptual and empirical nature. When phenomenal consciousness is the target of our theorizing activity, it is often the former that generate the most controversy. From the response to our target article, our connectionist vehicle theory is no exception in this regard. Our proposal identifies conscious experience with the brain's explicit representational vehicles in the form of stable patterns of neural activity. Commentators raise concerns about our reliance on *vehicles*, on *representation*, on *stable* patterns of activity, and on our *identity* claim. In this first part of our reply we address these concerns in that order.

1. Vehicles

A number of commentators (**Church, Dennett & Westbury, Kurthen, McDermott, Van Gulick**), including some who are sympathetic with the connectionist focus of the target article (**Cleeremans & Jimenez, Wolters & Phaf**), think that our exclusive focus on stable patterns of activity across the brain's neural networks is wrong. These patterns, they feel, might in some way be *necessary* for consciousness, but they surely can't be *sufficient*. They thus exhort us to augment our connectionist story with various kinds of computational processes in which these activation patterns are implicated. Wolters & Phaf, for example, suggest that they are the ingredients for "subsequent constructive processing", and it is these processes, not the patterns themselves, which are ultimately responsible for phenomenal experience. Cleeremans & Jimenez contend that patterns of activity are potentially available to consciousness, but whether they become so depends on a number of other factors, including "access by some other structure". And in a similar vein, Dennett & Westbury presume that it is their function in "modulating the

larger activities of the entire cortical meta-network" that mark these patterns for a role in phenomenal experience.

This objection strikes at the very heart of our connectionist theory of consciousness. In urging us to incorporate the computational roles in which neural activation patterns subsequently engage, these commentators are asking us to reject a vehicle theory, and adopt a process theory in its stead. We accept that connectionism has the resources to develop a process theory of consciousness. We also accept that some theorists will find this option irresistible, given the widespread presumption in favor of process theories. But the whole purpose of the target article is to present another option. The commentaries have made us realize, however, that it is not enough to observe that this part of the theoretical landscape is relatively unexplored. We need to explain why a vehicle theory of consciousness is attractive in its own right. Fortunately, this is relatively easy, because only a vehicle theory can satisfy one of the deepest intuitions we have about our conscious experience—that it makes a difference.

Perhaps the best way to see this is to consider the metaphysical implications of embracing a process theory of consciousness.[1] A process theory claims that the content of a representational vehicle is conscious when that vehicle has some privileged computational status, say, being available to an executive system, or being inferentially promiscuous. On this kind of story, consciousness is a result of the *rich* and *widespread* informational access relations possessed by a (relatively small) subset of the information bearing states of a cognitive system (see, e.g., **Cleeremans & Jimenez**, **Wolters & Phaf**).

There is, however, a certain amount of discord, among adherents of process theories, as to what *rich informational access* actually consists in. When philosophers and cognitive scientists talk of informational access, they often treat it as a notion to be unpacked in terms of the *capacity* of representational vehicles to have characteristic cognitive effects. This approach is evident, for example, in Block's characterisation of "access-consciousness", where he talks of representational vehicles being "poised" for use in reasoning, and the rational control of action and speech (1995, p.231). On this reading, what the process theorist asserts is that items of information are phenomenally conscious in virtue of the *availability* of their representational vehicles to guide reasoning and action. When Dennett, on the other hand, comes to explain phenomenal consciousness in terms of informational access relations, he has a distinctly different notion in mind (1991, 1993). Those contents are conscious, he claims, whose representational vehicles persevere long enough to achieve a persistent influence over ongoing cognitive events. This involves a somewhat different notion of informational access, since the focus has moved from the capacity of certain representational vehicles to guide reasoning and action, to their *achievements* in doing so. As a consequence, the flavor of Dennett's process theory of consciousness is different from most others found in the literature.

But can both of these interpretations of informational access be sustained by process theorists? We think not. It makes little sense to talk of a particular representational vehicle enjoying rich and widespread information processing *relations* in a cognitive system unless it is actually having rich and widespread information processing *effects*. Dennett, we believe, has seen this, and so avoids reading rich informational access in terms of the capacities of a select subset of representational vehicles. Instead, he concentrates on what these vehicles actually *do* in the brain—the impact they have on the brain's ongoing operations. As a result, phenomenal experience, according to Dennett, is like fame, a matter of having widespread effects. With regard to pain, for example, he argues that our phenomenal experience is not identifiable with some internal state which is poised to cause typical pain reactions in the system; rather; "it is the reactions that *compose* the 'introspectable property' and it is *through reacting* that one 'identifies'

---

[1] The following material is derived from O'Brien & Opie 1997, where we take up the motivation behind a vehicle theory at greater length.

or 'recognizes' the property" (1993, p.927). Consequently, in spite of the barrage of criticism that has been levelled at Dennett's account of phenomenal consciousness over the years, his position is actually more consistent with the general entailments of process theories.

But Dennett's work throws into sharp relief the deeply counter-intuitive consequences of adopting a process theory. To identify phenomenal experience with such information processing effects is to fly in the face of conventional wisdom. Consciousness, we intuitively think, makes a difference; it influences our subsequent cognitions and, ultimately, behavior. From a metaphysical point of view, this means that conscious experiences are, first and foremost, special kinds of *causes*: states that are distinct from and are causally responsible for the very kinds of cognitive effects that Dennett highlights. Consequently, cognitive scientists face a choice: either they must give up the idea that conscious states are causes, or they must give up on process theories of consciousness. Dennett exhorts these theorists to opt for the former, claiming that we shouldn't view it as ominous that such process theories are at odds with common wisdom: "On the contrary, we shouldn't expect a good theory of consciousness to make for comfortable reading... If there were any such theory to be had, we would surely have hit upon it by now" (1991, p.37). We think, on the other hand, that this is too high a price to pay. Our intuitions about the causal potency of our conscious experiences are some of the most deep-seated that we have. We give them up at our peril.

All of which brings us back to the motivation behind our vehicle theory of consciousness. To cleave to vehicles rather than processes is to hold that conscious experiences are *intrinsic* properties of the activity generated across the brain's neural networks. It entails that these experiences are determined independently of the cognitive causal roles in which these activation patterns subsequently engage. **Dennett & Westbury** think that such an approach is "confused": "If it turned out...that there was a subclass of stable patterns in the networks that did not play any discernible role in guiding or informing potential behavior, would their stability alone guarantee their status as part of phenomenal experience? Why?" Far from being confused, however, buying into a vehicle theory is the only way cognitive science can hope to do justice to one of our strongest intuitions about consciousness. It is only when phenomenal experience is an intrinsic property of the brain's representational vehicles that it can be a full-blooded cause of subsequent cognitive effects. It is the only way of making sense of the intuition that our behavior depends on our experience, not the reverse. Vehicle theories are thus very attractive in their own right. And only connectionism is in a position to exploit this fact.

Thomas & Atkinson and **Van Gulick** object to this last claim. They argue that we have failed to consider the possibility that among the classical vehicles of explicit representation there are distinct *types* (i.e., distinct subspecies of symbol structures), one of which might plausibly be aligned with phenomenal experience. What lends weight to this objection is the observation that both computer science and cognitive psychology appear to distinguish between different kinds of explicit representations. We find computer scientists referring to *simple* and *composite* data types, for example, and cognitive scientists referring to *frames*, *script*, *propositional encodings*, and *productions*, to name but a few. The suggestion is that the classicist might be able to exploit these distinctions to provide a *vehicle* criterion for consciousness. We don't think this will work. The distinctions between structured data types in computer science, and between the various kinds of representations in cognitive psychology, are based on the *differential computational roles* of these vehicles, rather than their intrinsic physical properties (see O'Brien and Opie forthcoming, for a detailed defense of this claim). To associate conscious experience with a particular type of explicit representation is, therefore, to adopt a process theory. Essentially the same point applies to **O'Rourke's** contention that a classicist could ground a vehicle theory in the distinction between explicitly represented rules (program) and explicitly represented information (data). This distinction is also grounded in computational relations, rather than intrinsic properties. Whether a sequence of tokens on a Turing machine's tape is an instruction or data is not

determined by the tokens themselves; it is determined by their computational effects on the machine's read/write head.

So we are thrown back to connectionism as the only plausible source of a vehicle theory. What this means, however, is that we must answer the question posed by **Dennett & Westbury**: If it isn't their causal effects, what is it about the patterns of activity across the brain's neural networks that makes them conscious? It is time to put some content into our vehicle theory of consciousness. We begin this process in the next section.

2. Representation

2.1 Why link consciousness and representation?

The connectionist theory of phenomenal experience proposed in the target article identifies consciousness with the brain's vehicles of explicit representation. Several commentators wondered about the general motivation for linking consciousness and representation in this way. **Perner & Dienes**, for example, observe that we spend a good deal of time re-assessing the *dissociation* studies in order to make room for proposal; but what we don't do, they think, is make out a strong case for *associating* representation and consciousness in the first place. This sentiment is echoed in various ways by **Ellis, O'Rourke**, and **Reeke**.

In the target article we suggest two quite different motivations for thinking that consciousness has got something to do with representation. We note, first, that what is special about cognitive science is its commitment to the computational theory of mind. The brain is thought to be in the business of representing and processing information, and cognition is understood in terms of disciplined operations over neurally realized representations. And we note, second, that from the first person perspective, phenomenal experiences would seem to carry information about either our own bodies or the world in which we are embedded. In this sense, conscious experiences are representational. These two motivations make it natural to seek a link between consciousness and representation. What is more, there would seem to be only two ways of doing this: either consciousness is to be explained in terms of the intrinsic properties of the brain's representational vehicles, or it is to be explained in terms of the computational processes defined over these vehicles. There just are no other available options if one wants both to explain conscious and to remain within the confines of cognitive science (something that tacitly is acknowledged by **Perner & Dienes**, when in their commentary they go on to defend a "Higher-Order-Thought" account of consciousness—a species of process theory).

2.2 Can vehicles be explicit representations?

In the case of the connectionist vehicle theory that we defend, stable activation patterns across the brain's neural networks are presumed to be the relevant vehicles of explicit representation. But a number of commentators have difficulties both with our reliance on network activation patterns as vehicles in this respect, and with our reliance on the notion of explicit representation more generally. The general thrust of their objections is that any theory that identifies consciousness with either stable network activity or the explicit representation of information in the brain will inevitably incorporate elements of computational process: a "pure" vehicle theory of the sort we propose isn't really a coherent option.

This charge is most straightforwardly developed by **Mac Aogáin** and **Wolters & Phaf**. They claim that network activation patterns cannot be treated as "free-standing" vehicles, either because they are the products of processes of relaxation (Wolters & Phaf) or because "there must be a process running in the background" to render them stable (Mac Aogáin). But we think both versions of this argument are invalid. While it is true that stable patterns of activity are generated and sustained by flurries of activation passing that spread across networks, the suggested conclusion (that these patterns themselves must in part be understood *as* processes)

doesn't follow. Network activation patterns are physical objects with intrinsic structural properties just as much as neurotransmitter molecules, neurons, and brains. That these latter entities need electrodynamic, biochemical and neurophysiological processes to generate and support them in no way undermines their status as spatiotemporally extended objects in their own right.

Church develops the argument in a different way. Her worry is that the distinction we draw between explicit and nonexplicit representation, especially as we develop it in the connectionist context, is ill-begotten: "neither the notion of coding in discrete objects nor the notion of active versus potentially active representation seems to help in specifying what is distinctive of [explicit] representation" However, Church's analysis is in one way incomplete, and in another quite mistaken. It is in part mistaken because nowhere do we unpack the distinction between explicit and nonexplicit in terms of "active versus potentially active representations". Here Church seems to have illicitly conflated "potentially explicit" with "potentially active". It is the potentially explicit information stored in a PDP network that governs its computational operations (target article, Section 4.2). Thus, potentially explicit information is not merely potentially active; it is active every time a network is exposed to an input. More importantly, though, Church's analysis is incomplete because explicit representation, on our account, requires more than that information be coded by physically discrete objects. It also requires that the physical resources used to encode each item of information be distinct from those used to encode others. Activation pattern representations satisfy this requirement; no network pattern of activity represents more than one distinct content. We might say that they are *semantically* discrete. Connection weight representations, on the other hand, fail to do so, because they encode information in a superpositional fashion; each connection weight contributes to the storage of many, if not all, of the stable activation patterns (explicit representations) the network is capable of generating.

At this point we should introduce **Clapin's** concerns about our representational taxonomy, as these specifically focus on the analysis of explicit representation that we have just employed. His assault has two prongs. The first maintains that in characterizing explicit representation in terms of semantically discrete parcels of information we differ in significant ways from the representational taxonomy on which ours is supposed to be based (viz., Dennett 1982). There is some justice in this charge. Dennett never states that for information to be represented explicitly, the physical resources used to encode each item of information must be distinct from those used to encode others (though in point of fact, all of the examples he invokes of explicit representation—sentences, maps, and diagrams—do have this character). But this is of no real concern. While Dennett's taxonomy provides us with a useful starting point, we are under no obligation to follow him in every detail. Our account, unlike Dennett's, is developed in the context of specific computational architectures (i.e., digital computers and PDP systems). Consequently, our taxonomy is more appropriately judged on the way it illuminates the different forms of information coding in these architectures. In this respect, we think the focus on semantic discreteness is actually essential. There is a crucial difference between the manner in which symbol structures and activation patterns on the one hand, and microcircuits and connection weights on the other, encode information. And this is what our distinction between explicit and nonexplicit forms of representation captures.

In the second prong of his assault **Clapin** insists that, contrary to our own analysis, activation patterns across PDP networks can and do participate in superpositional forms of information coding. He defends this claim by describing an imaginary three layer PDP which, because its input layer is divided into two subsets, is able to process two inputs simultaneously. He asserts that "in such a network the activation vector corresponding to the hidden units would superpositionally represent the two inputs". But here Clapin is using "superposition" merely to describe a process in which two (input) patterns are *combined* in some fashion to form a third (hidden layer) pattern. Superpositional *representation* is a quite different notion. As we have

already noted, a computational device represents information in a superpositional fashion when the physical resources it employs to encode one item of information overlap with those used to encode others. It is standard practice in PDP modelling to suppose that the content of an activation pattern representation is fixed by the point it occupies in the network's "representational landscape". This landscape, which can be revealed by such numerical techniques as cluster analysis and principal components analysis, is constituted by the range of stable activation patterns that a trained-up network can generate in response to its full complement of possible inputs. Such a story about representational content seems to rule out superpositional coding in activation patterns, as it is impossible for an individual activation pattern to occupy, at the one time, two or more different points in this representational landscape. (For a more detailed discussion of the issues raised in this paragraph, see Clapin & O'Brien 1998.)

Whereas **Clapin** accuses us of illicitly augmenting Dennett's taxonomy of representational styles, **Dennett & Westbury** accuse us of overlooking an important further taxon: *transient tacit representations*. These are "tacit representations which are available for a system's use only when that system is in a particular state". Dennett & Westbury claim that the "stable connectionist patterns championed by O'Brien and Opie are presumably just such sorts of mental representations". They go on to suggest that we ignore the possibility of tacit representations that are not hardwired. We are quite perplexed by this charge, for two reasons. First, while activation pattern representations are certainly transient features of PDP networks, they do not appear to satisfy the general requirements of tacit representation as it is characterized in Dennett 1982. Information is represented tacitly, for Dennett, when it is embodied in the primitive operations of a computational system; it is the means by which a system's basic know-how is implemented (p.218). It is hard to see how stable activation patterns can fulfil this role, given that such patterns are themselves dependent on a network's connection weights. Surely the latter, not the former, embody the primitive computational know-how of a PDP system. Second, we don't claim that connectionist tacit representations are "hardwired". Quite the reverse, in fact (see Section 4.1). It is only in virtue of the plasticity of tacit representation—through modifications to a network's connection weights—that PDP devices learn to compute. That tacit representations are not hardwired is thus one of the fundamental commitments of connectionism.

Schröder also upbraids us for overlooking something in Dennett's taxonomy. Our account of explicit representation, he observes, is based on purely structural criteria; yet Dennett's combines both structural and functional elements. What is more, Schröder continues, Dennett's characterisation is more consistent with a recent influential discussion by Kirsh, according to whom explicitness really concerns "how quickly information can be accessed, retrieved, or in some other manner put to use". Explicitness, Kirsh concludes, "has more to do with what is present in a process sense, than with what is present in a structural sense" (1990, p.361). On this analysis, it is impossible to develop an adequate characterisation of explicit representation without invoking process criteria. This is very bad news for anyone who wants to develop a vehicle theory of consciousness, since it seems to suggest that the project is misguided right from the start.

We are tempted to respond here by saying that this is nothing more than a terminological dispute; that two different but equally legitimate conceptions of "explicit representation" are available in cognitive science—a structural conception, and a process conception—and the squabble is over who gets to use this term. Unfortunately, the issues here are much murkier than that. Like consciousness, representation is one of the knottiest problems in contemporary philosophy of mind. What really underpins this dispute are two profoundly different ways of thinking about mental representation.

It is fashionable in cognitive science and philosophy of mind to suppose that the mind's representational content must be unpacked in terms of causal transactions between the brain

and the environment in which it is embedded. On this view content has got very little to do with the intrinsic properties of the brain's representational vehicles, and everything to do with their (e.g., actual, counterfactual or historical) causal relations with the world. It is for precisely this reason that **Lloyd** counsels us to dissociate phenomenal content from representational content. Since a vehicle theory holds that "states of consciousness are identical with states individuated by their intrinsic properties, rather than by their functional context", he writes, it implies that conscious experiences "cannot carry information about an outside world, and so cannot represent an outside world (at least by most accounts of representation)". Lloyd's suggestion is that on the approach we are taking, an approach with which he is highly sympathetic, consciousness should not be viewed as a kind of representational vehicle, it should simply be understood as "a complex state".

In the current climate in cognitive science and the philosophy of mind, this is good advice. Given the standard take on representational content, we cannot even begin to formulate a vehicle theory that identifies consciousness with explicit representation—any identification of consciousness with representational content will entail a process theory. But rather than redescribing our connectionist vehicle theory, we think it is the standard analysis of representational content that ought to come under pressure. For while it is commonplace nowadays to hear that connectionism radically alters our conception of human cognition, what has yet to be fully appreciated is how radically it alters our understanding of way the brain goes about representing the world. This point deserves a subsection of its own.

2.3 Connectionism and representational content

The standard analysis of representational content is, in large measure, a legacy of classical cognitive science. Given that digital computations inherit their semantic coherence from rules that are quite distinct from the structural properties of the symbols they apply to, classicism appears to place few constraints on a theory of mental content. Thus the causal, functional and teleofunctional theories that dominate the current literature are all, prima facie, compatible with the idea that mental representations are symbols. All of this changes, however, when we move across to connectionism. Given its foundation in the PDP computational framework, connectionism undermines the distinction between representational vehicles and the processes that act on them. A PDP system is not governed by rules that are distinct from the structural properties of its representational states. Instead, it computes by exploiting a *structural isomorphism* between its physical substrate and its target domain.

Consider, as an example, NETtalk, probably the most talked about PDP model in the connectionist literature (Sejnowski & Rosenberg 1987). NETtalk transforms English graphemes into appropriate phonemes, given the context of the words in which they appear. The task domain, in this case, is quite abstract, comprising the (contextually nuanced) letter-to-sound correspondences that exist in the English language. Back propagation is used to shape NETtalk's activation landscape—which comprises all the potential patterns of activity across its 80 hidden units—until the network performs accurately. Once it is trained up in this fashion, there is a systematic relationship between the network's activation landscape and the target domain, such that variations in patterns of activation systematically mirror variations in letter-to-sound correspondences. It is this structural isomorphism that is revealed in the now familiar cluster analysis which Sejnowski and Rosenberg applied to NETtalk. And it is this isomorphism that makes it right and proper to talk, as everyone does, of NETtalk's having a *semantic metric*, such that its activation landscape becomes a *representational* landscape. Furthermore, and most importantly, it is this isomorphism that provides NETtalk with its computational power: when NETtalk is exposed to an array of graphemes, the structural isomorphism dispositionally embodied in its connection weights automatically produces the contextually appropriate phonemic output.

Because it is grounded in PDP, and because PDP computation requires the presence of structural isomorphisms between network patterns and their target domains, connectionism brings with it not just a different way of thinking about human cognition, but a profoundly different way of thinking about the content of mental representations. Instead of thinking in terms of *causal transactions* between the brain and its embedding environment, we are required to think of representational content as a special kind of *correspondence* between intrinsic properties of neural activation patterns and aspects of the world. A few of years ago this might have seemed a serious objection to connectionism, as this general conception of representational content, which has a venerable history in philosophy, was thought to suffer from a number of serious flaws (see Cummins, 1989, chp.3). But recently a number of theorists have started to take this approach very seriously.[2]

Of course any talk of a structural isomorphism *theory* of representational content is clearly premature. We have merely suggested a direction in which connectionists might head in their efforts to tell a more complete story of human cognition. Nevertheless, this very different way of thinking about mental content, because it focuses on the structure of the brain's representational vehicles rather than their casual relations, complements the vehicle theory of consciousness we have proposed. According to the latter, conscious experience is the brain's explicit representation of information in the form of neural activation patterns. According to the former, these activation patterns are contentful in virtue of relations of structural isomorphism between them and features of the world. We will see, in the sections to follow, that the marriage of these two ideas offers some prospect of a perspicuous nexus between the material properties of the brain and the phenomenal properties of our experiences.

3. Stability

3.1 Why stable activation patterns?

Lloyd, **Mangan**, and **Reeke**, all of whom express some sympathy for our proposal, wonder why we have restricted our account of consciousness to *stable* patterns of neural activity. This, they think, both prevents us from taking advantage of the "dynamic" properties of neural activity, and makes it impossible for our vehicle theory to explain the "flights" of consciousness—its fluidity and evanescence. With regard to the former, Reeke thinks it unfortunate that by focusing on stable patterns we have neglected the role that the complex reentrant interaction between neural patterns has on the emergence of phenomenal experience across large portions of the brain, and Mangan is aghast that we have needlessly ignored the contribution that stabilizing networks can make to conscious experience. With respect to the latter, Reeke questions how stability "even in suitably quantized chunks of time" can explain the smooth flow of experience, while Lloyd suggests that it is more likely that instability accounts for the unmemorability and nonreportability of certain kinds of conscious episodes.

These worries are important. But as we observe in the target article (Section 5.1), there is a straightforward reason for making stability such a central feature of our connectionist account. A vehicle theory identifies consciousness with the brain's explicit representation of information, and only stable patterns of activation are capable of encoding information in an explicit fashion in PDP systems. Patterns of activation across PDP networks are contentful, we argued in the previous section, in virtue of being structurally isomorphic with certain aspects of a target domain. But such structural isomorphs cannot be realized in PDP networks unless they achieve

---

[2] Two theorists who have kept the torch of structural isomorphism burning over the years are Palmer (1978) and Shepard (Shepard and Chipman, 1970; Shepard and Metzler, 1971). But more recently, Blachowicz (1997), Cummins (1996), Edelman (forthcoming), Files (1996), Gardenfors (1996), and Swoyer (1991), have all explored, though in different ways, the idea that relations of isomorphism might ground the content of the brain's representational vehicles. For a general discussion of these issues, see O'Brien forthcoming.

stable patterns of activity. This is because prior to stabilization, there are no physically objects present in these networks whose intrinsic structural properties can stand in this kind of relation with elements of the target domain. A connectionist vehicle theory must, therefore, identify phenomenal experience with stable patterns of activity across the brain's neural networks.

What is more, contrary to what these commentators claim, a vehicle theory that focuses on stable patterns can both exploit the dynamical properties of neural activity and explain the fluidity and evanescence of conscious experiences. Against **Reeke**, for example, the connectionist vehicle theory we are proposing doesn't neglect the complex interplay between largescale patterns of activity across the brain. It is precisely this interplay, we think, that is responsible for the inter-network processes that bind phenomenal elements together to construct the unified cognitive subject (see Section 7 below). Moreover, given the time scale at which network stabilizations can occur in the brain, it is not implausible to suppose that the "seamless" flow of experience is actually composed of quantized phenomenal elements, each equipped with its own distinct representational content. In this sense, the "flights" of consciousness that **Lloyd** (following James 1890) highlights, are plausibly reconceived as rapid sequences of "perchings", and it is their rapidity, not the absence of stability, that accounts for the unmemorability and nonreportability of certain kinds of conscious experiences.

3.2 What is a stable activation pattern?

We have just argued that a connectionist vehicle theorist is committed to identifying consciousness with stable patterns of activity across the brain's neural networks. According to a number of commentators, however, this just leads us into further hot water. For them, and here we have in mind **Cleeremans & Jimenez**, **Dennett & Westbury**, **Gilman**, **Lloyd**, **Pólya & Tarnay**, **Schröder**, and **Taylor**, our characterisation of stability is problematic. In the target article we opt for a simple story according to which a neural network realizes a stable pattern when its constituent neurons are firing simultaneously at a constant rate (Section 5.1). Several commentators observe that such a definition is incomplete absent a relevant time scale. We accept this criticism. But this problem is not ours alone: connectionist theorizing about cognition in general is deeply committed to stability.

The failure to properly distinguish between the properties of digital simulations of PDP networks, and their real counterparts, makes it possible to miss the significance of stability. In simulations, a neural network's activation pattern is modelled as an array of *numerical activation values*. These activation values are numerical descriptions of the spiking frequencies of real neurons. They are periodically updated by the algorithms that model the network's activity. In such a simulation, processing proceeds via a sequence of activation patterns, as the network relaxes into a solution. And this gives the impression that prior to stabilization, a neural network jumps between specific points in its activation space. But this picture is misleading. Whenever one employs a numerical value to describe a continuously variable physical property, one is imposing on this property an instantaneous value. This is fine for many properties, such as charge, or velocity, which possess a determinate value at every instant (although, in the case of velocity, one relies on the assumption that space and time are themselves continuous). But a spiking frequency, or firing rate, is importantly different. Because neural spikes are discrete events, neural spiking rates *do not have instantaneous values*; the notion of a rate, in this case, only makes sense relative to a time scale. What this means is that digital simulations of PDP systems contain an important idealization: at each tick of the time clock, as the model network settles towards a response, constituent units have their firing rates adjusted from one instantaneous value to another. In a real network, by contrast, stabilization is a continuously unfolding process which sees constituent neurons adjust the absolute timing of their spikes until a determinate firing rate is achieved. Prior to stabilization, neural networks don't jump around between points in activation space. Stabilization is the process by which a network first *arrives* at a point in activation space, and hence takes on a determinate activation pattern.

Developing a satisfactory characterization of stability is therefore a task in the theoretical foundations of connectionism. Its solution will depend on the computational significance that attends the precise temporal properties of neuronal spiking trains, an area of neuroscience, as **Gilman** points out, where there are a number of significant and unresolved questions. Given the chemical dynamics of neural networks, however, it is reasonable to conjecture that the time scale relevant for stability is in the order of tens of milliseconds (see Churchland & Sejnowski 1992, Chp.2).

Dennett & Westbury have a further problem with of our account of stability. They claim that "it is easy to imagine a network sampling a number of points from another network and finding them stable because of its (the sampler's) characteristics, even though there is nothing in the sampled state which shows the stability". This, they contend, indicates that there may be forms of stability in the brain that are not purely intrinsic to individual neural networks: "Stability is as much a function of the sampler as of the sampled". What's the worry here? We suppose it's that certain kinds of meaningful network interaction may occur in the brain in the absence of intrinsic stability, and this puts some pressure on our claim that intra-network stability plays an important inter-network information processing role (that stability begets stability—see target article, Section 5.1). It is reasonable to wonder about the potency of this threat, however. Our conjecture about the role of network stability at least has the merit of being based on one of the assumptions of connectionist theorizing: that downstream networks cannot complete their processing cycles (and thereby generate explicit information) unless their inputs from upstream networks are sufficiently stable. In the absence of an alternative neurocomputational explanation of inter-network information processing, we think the worry raised by Dennett & Westbury is idle.

3.3 Stability in simulations and artificial networks

We will finish this section by examining a number of worries about the *absence* of stable patterns in digital simulations of PDP networks, and their *presence* in various kinds of artificial and in vitro networks. **Cleeremans & Jimenez** and **Dennett & Westbury** are baffled by our claim (made in Section 5.1) that stable activation patterns, considered as complex physical objects, are absent in digital simulations of PDP systems. In the target article we defend this claim by arguing that an activation pattern across a real network has a range of complex structural properties (and consequent causal powers) that are not reproduced by the data structures employed in simulations. Cleeremans & Jimenez reproach us for borrowing Searle's (1992) "mysterious" notion of the causal powers of biological systems. However, far from being mysterious, our appeal to "causal powers" is extremely prosaic. One of the uncontroversial properties of a wing is that it can generate lift. We have yet to come across a digital simulation of aerodynamic phenomena that has this capacity. Real wings have causal powers not possessed by their simulated counterparts. Similarly, in digital simulations of PDP networks, the data structures (i.e., the numerical arrays) that represent network activation patterns, do not have the same causal powers as those patterns.

Dennett & Westbury disagree. They claim that the stability of a virtual machine is every bit as powerful as the stability of an actual machine. However, we think this is demonstrably false for the reasons we develop in the target article: there are vast temporal asymmetries between real PDP networks and their digital simulations. These temporal asymmetries arise, because the structural properties (and hence causal powers) of a numerical array in a digital computer are quite different from those possessed by the web of active elements (however these might be physically realized) that make up a real PDP network. Most obviously, there are causal connections between the elements of a PDP network that simply don't exist between the variations in voltage that physically implement a numerical data structure. The manner in which activation patterns evolve in each case is thus quite different. In a real network, this evolution is dynamical: all the connected elements have their activity continuously modulated by the activity

of others, until the network settles into a stable state. The activation pattern in a simulated network, by complete contrast, evolves through the operation of an algorithm which updates activation values individually.

On the other side of this ledger we find a group of commentators wondering about the presence of stable patterns of activity in various kinds of artificial PDP networks. **McDermott**, for example, notes that while we deny conscious experience to a digital simulation of such a network, we "don't quite say whether a network of digital computers, each simulating a neuron of the classic linear-weighted-sigmoid-output variety, would be conscious", and hence suspects that our intuitions are inconsistent. And in a similar fashion, **Perner & Dienes** claim that the trouble with our vehicle theory is that "it would be easy to set up a real PDP network made up of electronic chips with a stable pattern of activation", something, they think, that would have "no more consciousness than a thermostat" (see also **Gilman**). This is a moment for bullet biting. Our connectionist proposal is that conscious experience is the explicit representation of information in the form of stable activation patterns across neurally realized PDP networks. However, we accept that not all the properties of neural networks are necessary for the physical implementation of these stable patterns. As we remark in the target article (Section 4.1), connectionism is founded on the conjecture that PDP isolates the computationally salient properties of neural networks, despite ignoring their fine-grained neurochemistry. In principle, therefore, we can envisage artificial PDP networks replete with all the intrinsic structural properties and consequent casual powers that matter for computation in the brain (though whether one could do this with a network of digital computers, or a network of electronic chips, is something about which we remain agnostic). In such case, our proposal commits us to ascribing the same elements of phenomenal experience to these artificial networks as we do to their neural network counterparts in our heads.

Stated this baldly, we know that many readers (including **Perner & Dienes** presumably) will baulk at these implications of the vehicle theory we are defending. Do we really want to be committed to an account of consciousness that ascribes experiences to artificial networks? **Pólya & Tarnay** develop this worry in a particularly vivid way by noting that there is a sorites paradox looming here: Just how complex does a network pattern have to be before it is capable of conscious experiences? And **Gilman** simply asks: "Is a 28 cell network, stable in vitro, concious? If so, of what?". Vehicle theorists are not completely without resources in this regard. They can appeal to the theory of representational content we briefly sketched above (Section 2.2) to provide some minimal constraints on both the emergence of conscious experiences and the nature of their phenomenal properties. But there is a deeper issue here. The air of implausibility that surrounds our connectionist proposal at this juncture is one that envelopes *all* materialist theories of consciousness. Whatever physical or functional property of the brain's neural networks one cares to name, one will always be vulnerable to these kinds of considerations. For example, process theorists who explain consciousness in terms of the rich and widespread processing relations enjoyed by a (relatively small) subset of the brain's representational vehicles have the problem of specifying just *how* rich and widespread these processing relations must be. Sorites considerations are, then, just another way of highlighting the much vaunted explanatory gap with which all materialists must contend. Towards the end of the target article we suggest how our connectionist proposal might close this gap (Section 5.4). Not surprisingly a number of commentators are not convinced. We address their concerns in the next section.

4. Identity (Or: Once More Into The Explanatory Gap)

Suppose on some future day in the golden age of connectionist neuroscience we have compiled an exhaustive account of all the kinds of activation patterns that are generated in all the different neural networks in our heads. And suppose further that experimental research demonstrates that there is a precise one-to-one mapping between these kinds of activation patterns and

specific types of phenomenal experience. Would this show that conscious experiences are *identical* to stable patterns of network activity in the brain? **Ellis**, **van Heuveln & Dietrich**, **Kurthen**, **Newton**, and **Velmans** think not. Our connectionist vehicle theory, they charge, whatever its virtues as an account of the neural *correlates* of consciousness, doesn't have the resources to bridge the explanatory gap.

van Heuveln & Dietrich offer two reasons for this. First, although you might one day observe stable activation patterns in my brain, you will never observe my phenomenal experiences. Second, we can conceive of creatures that have stable patterns of activation but no phenomenal experiences. The conclusion, in both cases, is that stable activation patterns cannot be identical to phenomenal experiences (see also **Kurthen**). Neither argument is terribly convincing. With a story in place that exhaustively matches network activation patterns with specific phenomenal experiences, we think it would seem quite natural to think that in observing the former we are observing the latter. Of course, in *observing* the former, we would not *have* the latter, but this is quite another matter (try substituting 'have' for 'observe' in both of the premises of van Heuveln & Dietrich's first argument). As for the second argument, we give some reasons in the target article for not placing too much faith in our conceptual powers. With **Newton**, we think conceivability arguments are incapable of yielding any substantive conclusions regarding matters of ontology and metaphysics.

Kurthen thinks the explanatory gap would persist for a different reason. "Even if the state of unconsciousness would be inconceivable in the face of the mechanism [i.e., the activation patterns] in question", he writes, "the co-occurrence of...phenomenal consciousness and the analogically structured mechanisms...could neither explain the internal constitution of the relata themselves..., nor the kind of their relationship". In a similar vein, **Ellis** and **Newton**, while granting that our vehicle theory can do much to close the explanatory gap, argue that the kind of "perspicuous nexus" required for a satisfactory reductive explanation of consciousness has not been achieved and is perhaps unachievable. For Ellis, our theory hasn't explained why the information explicitly represented in the brain is conscious, when information explicitly represented elsewhere—on this sheet of paper, for example—isn't. For Newton, the problem concerns the different stances that are involved in being a conscious subject, and in scientifically observing an active brain. According to Newton, while both these stances are made possible by the physical components of brains, passing between them entails a gestalt shift, consequently, "no single coherent (nondisjunctive) description will capture the aspects of *both* stances".

These observations are well taken. We accept that the mere co-occurrence of activation patterns and experience would be insufficient to ground an identity claim. And we accept that there are special problems in attempting to develop an intelligible connection between the micro-mechansims of the brain and the macro-properties of consciousness. But we have already sketched (in bare outline) the kind of resources that might provide the required perspicuous nexus. We are referring to the theory of representational content according to which the "internal structure" of conscious experiences is determined by relations of structural isomorphism between network activation patterns and certain properties of the world. Of course, this is no more than a hint of a suggestion (and would require a detailed research programme to even begin to do it justice), but the marriage of a structural isomorphism theory of mental content with a vehicle theory of consciousness does offer some prospect of closing the explanatory gap.

Suppose, then, that in addition to showing that there is a one-to-one mapping between activation patterns and types of phenomenal experience, our future neuroscience also reveals that in every case there is a structural isomorphism between these patterns and the properties of the world represented in the corresponding phenomenal experiences. Would this be enough to close the door on consciousness? **Velmans** thinks not and develops what is perhaps the most familiar line of reasoning that maintains a distance between our hypothesis about the neurocomputational substrate of consciousness and its phenomenal properties. "One

might…know everything there is know about the 'shape' and 'dimensionality' of a given neural activation space", he writes, "and still know nothing about what it is like to have the corresponding experience". Indeed, anticipating our earlier discussion of artificial networks, Velmans, with a Nagelian twist (1974), asks us to suppose that we arrange "a net to operate in a nonhuman configuration, with an 'activation space shape' which is quite unlike that of the five main, human, sensory modalities". According to Velmans, we cannot know what such an artificial network would experience. And "if we can know the 'shape' of the space very precisely and still not know what it is like to have the experience, then having a particular activation [pattern] can't be *all there is to having an experience*".

Now, initially, there is a very straightforward response to this line of reasoning. Surely, a good appreciation of the topology of a network's representational landscape would furnish a good deal of information about what it is like to occupy one of its points. Once we have the structural isomorphism theory of representational content in the foreground, however, another more speculative and altogether more intriguing response becomes possible. Just for bit of light relief, therefore, we'll finish this first half of our reply by sketching it.

Velmans's argument, the form of which has generated an enormous amount of philosophical discussion, seeks to derive an ontological conclusion (that phenomenal experiences are *not identical* to activation patterns) from a purported epistemic asymmetry (that we can *know* all there is know about the latter but *not know* the former). The standard materialist riposte asserts that this epistemic asymmetry doesn't entail a metaphysical gap between mind and brain, but merely highlights different ways in which the same phenomenon (our phenomenal experiences) can be known (see, e.g., Lewis 1990, Nemirow 1990, Churchland, 1990, 1995). But we think this reply doesn't get it quite right. The problem isn't that the argument equivocates between different *ways* of knowing, it's that it makes a mistaken assumption about what *knowing* is in the first place.

Regardless of whatever else might be necessary, knowledge implicates representation. On the theory of representation that we are considering, this requires the presence of structural isomorphisms between patterns of neural activity and certain aspects of the thing known. But now consider what is required to know about phenomenal experiences (rather than about aspects of the world by having phenomenal experiences). Since phenomenal experiences are activation patterns across neural networks, knowing about them requires one to generate structural isomorphs of these patterns. But now something very interesting happens. To have *exhaustive* knowledge about a phenomenal experience, on this analysis, one must generate a neural activation pattern that is exactly structurally isomorphic with the activation pattern one seeks to know—a pattern which in turn is isomorphic with some aspect of the world. Transitivity of structural isomorphism thus dictates that in order to have complete knowledge about a phenomenal experience, one must *reproduce* this very experience in one's head.

There is nothing mysterious about this. It would seem to be a fairly straightforward consequence of marrying the two materialist theories we have been considering: the vehicle theory of consciousness and the structural isomorphism account of representational content. But it is a result that completely unravels **Velmans's** argument and its variants wherever they appear in the literature. If Mary, the colorblind neuroscientist in Jackson's famous thought experiment (1982), knows all the "physical information" about red experiences prior to leaving her black and white room, then her brain must have reproduced the relevant activation patterns, and hence the experiences. On the other hand, if she doesn't know what it is like to have a red experience prior to leaving her room, then she doesn't know all the physical information. Either way, the argument will no longer deliver up its familiar conclusion.

Part II: Empirical Plausibility

Whatever the conceptual merit of a vehicle theory of consciousness, such an account ultimately stands or falls on the empirical evidence. Many commentators, some of whom express sympathy with our project, nevertheless feel that the evidence is against us. Their concerns range from the general plausibility of a vehicle theory of consciousness to our specific attempts to deal with what we term the dissociation studies (target article, Section 2). In this second part of our reply we address these worries. We then finish by considering whether the vehicle theory we have defended can go beyond the mere "what" of consciousness, as one commentator (**Carlson)** so eloquently puts it, to tell a coherent story about the "who" of consciousness: the active, unified subject to whom conscious experiences belong.

5. General Empirical Concerns

5.1 Consciousness is limited, while the unconscious is vast

A very general worry raised by **O'Rourke** is that our picture of unconscious mental activity is seriously awry, because it inverts the usual picture, championed by Baars (1988, 1994), of the relationship between the conscious and the unconscious: the former is limited, serial and slow; the latter vast, parallel and speedy. O'Rourke thinks that by treating unconscious activity as relaxation processes, thereby excluding the possibility of unconscious explicit representations, we seriously limit the role of the unconscious. Consequently, we aren't in a position to offer serious candidate explanations for even simple cognitive phenomena such as recall.

We'll take up the issue of the explanatory resources available to a connectionist vehicle theorist below, but let us first indicate where we differ with **O'Rourke's** assessment. In the target article we make much of the ways in which connectionist representation and processing differ from their classical counterparts (Section 4). In particular, although a classicist is committed to a great many unconscious, explicit representations in order to explain cognition, connectionists can dispense with these, because they have available far richer nonexplicit representational resources. What makes the difference is the fact that in PDP systems information storage and information processing depend on a common substrate of connection weights and connections. Since this information storage is superpositional in nature, the processing in a PDP system is causally holistic: *all* of the information nonexplicitly encoded in a network is causally active *whenever* that network responds to an input. Thus, along with a revolution in our understanding of consciousness, the connectionist vehicle theory brings with it a quite different way of thinking about the causal role of the unconscious. It entails that unconsciously represented information is never causally active in a functionally discrete fashion. This is not to say that all information processing in the brain has this flavor; conscious contents, precisely because they are explicitly represented, are causally discrete. But unconscious information processing, according to the vehicle theory, is always casually holistic.

So we accept **O'Rourke's** claim that the unconscious aspect of mental activity is vast. Where we differ is in how we picture that activity. O'Rourke imagines that unconscious processes are defined over a vast number of explicit representations. We take unconscious processes to involve the causally holistic operation of all the nonexplicit information stored in the weights and connections of neurally realized PDP networks. By comparison with the contents of consciousness the extent of this information is certainly vast, but such information does not take the form of stable patterns of activation in neural networks, so it is entirely nonexplicit.

Incidentally, this last is something that **Velmans** disputes. He rightly observes that information unconsciously represented in memory is constantly "causally active in determining our expectations and interactions with the world", and thinks that it must, therefore, be explicitly encoded. It's hard to see how such a view could be sustained. First, it is contrary to the orthodox

understanding of both digital and PDP systems, whereby tacitly (and hence, nonexplicitly) represented information plays a pivotal computational role. Second, and perhaps more importantly, it is precisely because classicism is committed to a vast amount of unconscious, causally discrete information processing that the infamous "frame problem" is so acute for this approach to cognition. The connectionist vehicle theory, with its promise of a casually holistic unconscious, appears ideally placed to provide a more realistic solution.

5.2 Aren't There Too Many Stable Activation Patterns?

It is at this point that a very common objection comes to the fore. Aren't there simply too many stable patterns of activation in the brain for us to seriously pursue a connectionist vehicle theory of consciousness? This objection is raised, in one form or another, by **Cleeremans & Jiménez**, **Gilman**, **Mangan**, **Perner & Dienes**, and **Van Gulick**.

Mangan claims that our account is "at odds with virtually all existing PDP models of neural activity". He takes it to be a fundamental feature of connectionist theorizing that there are lots of relaxed PDP networks in the brain that don't generate any conscious experience. This is a curious claim. Connectionist theorizing about phenomenal consciousness is in its infancy. Mangan gives the impression that we are swimming against a great tide of connectionist thinking in this area, but we think it is premature to start looking for a consensus. Apart from some suggestive comments in McClelland &Rumelhart 1986, Smolensky 1988, and Mangan 1993b, and the pioneering work of Lloyd (1991, 1995, 1996), there hasn't really been much work on the foundations of a distinctively connectionist approach to consciousness. The ultimate shape of such a theory is still very much up for grabs. It is in the spirit of exploring largely uncharted territory that we make our suggestions.

All the listed commentators are concerned about the existence of stable patterns of activation in the brain which don't contribute to consciousness. Suggested locations for such activation patterns are: the retina, the lateral geniculate nuclei, the sites of early visual processing in general, the spinal cord, and the sites of both physical and psychological reflexes. These suggestions require separate treatment.

Gilman is worried that fast automatic mechanisms, which aren't typically thought to give rise to conscious experiences, "may be excellent exemplars of consistently behaving networks". That is, the networks responsible for early perceptual processing, and for both physical and psychological reflexes, may enter stable (if transient) states of activation. However, when it comes to physical reflexes, we wonder whether it is not more plausible to suppose that such mechanisms involve either: (1) networks which connect input to output in a single processing cycle, with no stable intermediaries (we have in mind here simple reflexes such as withdrawal reflexes); or (2) in the case of autonomic reflexes (such as those involved in the maintenance of respiration), networks which settle on limit cycles in activation space, rather than point attractors, and hence never achieve stable activation.

When it comes to perception, and in particular the early processing of visual, auditory and speech signals, classical accounts presuppose a great deal of unconscious, explicitly represented information. Perceptual processing is assumed to be hierarchical in nature, beginning with a first stage of representations that are transduced from environmental input, transformations of which lead to further interlevels of explicit representation.[3] The contents of sensory consciousness correspond with some privileged stage in the processing of input[4], but the vast majority of the explicit representations generated during input processing are taken to be unconscious. **Gilman**

---

[3] Marr's (1982) theory of visual representation and processing is the archetypal account.

[4] Fodor, for example, suggests that we identify them with the *final* representations of input processing. These are the representations "most abstractly related to transduced representations" (1983, p.60). Jackendoff (1987), on the other hand, argues that it is intermediate representations whose contents are conscious.

and **Van Gulick** seem to be persuaded by these default (process model) assumptions. However, we don't think they are compulsory for a connectionist. The analysis we presented in our target article (Section 5.2) suggests that phenomenal consciousness is exceedingly rich; far richer than classically-inspired theorists are generally willing to acknowledge. Moment by moment perceptual experience, in particular, embraces a multitude of object features and properties, at many levels of detail. In other words, there is actually a great deal of phenomenology to play with when framing connectionist theories of perception. Consequently, connectionists are in a position to advance PDP models of, say, vision, that posit a raft of stable activation patterns, without in any way undermining the connectionist vehicle theory of consciousness.

However, **Gilman** and **Van Gulick** find it just implausible to suppose that all the various stable patterns of activation that arguably emerge in early perception should feature as elements of visual experience. But this looks more like an article of faith than a well supported empirical conjecture. Zeki, for one, argues that all areas of the cerebral visual cortex, including V1 (the cortical region at the lowest level in the processing hierarchy) contribute to visual experience (1993, p.301). It strikes us as more than reasonable that the contents of those explicit representations implicated in early perception—the boundaries and edges of vision, the phonemes and phrases of linguistic input—are the very elements of our sensory experience. Lloyd concurs. With regard to the "lesser sensations" of vision, the low-level visual representation of edges, figural boundaries etc., he claims:

> These are as much a part of our conscious perception of any scene as the high-level awareness of the names and meanings of things. Our awareness of them is fleeting and vague, but real and easily intensified with a shift of attention. (1991, p.454)

This plausible view sits comfortably with the connectionist account according to which each element of consciousness corresponds to the stable activation of a neural network. Sensory experience is simply the sum total of all the explicit representations that are generated during input processing.

But what of stable activation patterns in the retina, in the LGN, or, for that matter, in the spinal cord? Again, we suggest, it is simply an article of faith to reject such structures as legitimate sites of phenomenal experience. What grounds do we have for drawing a line somewhere in the brain, with conscious contents on one side, and unconscious contents on the other?[5] We simply don't know enough about the way informational contents are fixed in the brain (single cell recordings are of only limited help here) to categorically reject stable patterns of activation in the retina, or even the spinal cord, as components of conscious experience. Once one allows that the elements of phenomenal experience are generated at multiple discrete sites scattered throughout the brain (as suggested by deficit studies), then it becomes very difficult to motivate a boundary drawn anywhere within the CNS.

Regarding what Gilman calls psychological reflexes (he seems to have in mind here some of the mechanisms responsible for, say, speech perception, or higher thought processes), we think connectionism is in a position to suggest plausible accounts of these phenomena that dispense with stable, but unconscious intermediaries. Single-step psychological mechanisms connecting, say, a perceptual input (e.g., a word) with a response of some kind (e.g., an associated word, an anagram, etc.) can potentially be treated as relaxation processes across single or multiple networks. We explore this idea in more detail in the next section.

---

[5] A process theorist could, presumably, come up with some sort of functional criterion here, as Van Gulick suggests, but this begs the question against a vehicle theory, which rejects the claim that such criteria are constitutive of phenomenal experience.

5.3 Non-contrastive Analyses

If the worries discussed in the previous section are less than devastating, they do at least flag the crucial difficulty for a defender of the connectionist vehicle theory of consciousness: How, in the light of existing studies, does one justify the identification of phenomenal experience with the explicit representation of information in the brain? This problem has two parts, which line up with Dulany's useful distinction between *contrastive* and *non-contrastive* analyses (1991, pp.107-11).

A contrastive analysis is one that makes differential predictions explicitly designed to test for the presence of unconscious, explicit representations. The dissociation studies (target article, Section 2) are of this type. The first problem facing a vehicle theorist is to account for the weight of relatively direct evidence, provided by contrastive analyses, for the dissociation of phenomenal experience and explicit representation. We revisit this issue in the next section.

A non-contrastive analysis simply takes the dissociation of phenomenal experience and explicit representation for granted. Most speculative theories in cognitive psychology belong to this category (e.g., typical theories of memory, of learning, of perception, and so on). Insofar as they are successful, such analyses provide *indirect* support for the existence of unconscious, explicit representations, by way of a (presumed) inference to the best explanation. Thus, the second problem facing a vehicle theorist is to provide alternative explanations for those phenomena that have been successfully treated on the standard model. This is no small undertaking (!), but since several commentators focus on this issue, we will address it briefly here (and also in our remarks about implicit learning and subliminal priming in the next section).

Most explicit about this issue is **Schröder**, who claims that:

> Every successful theory about a cognitive capacity implies (under a realist conception of science) that the entities which are postulated by the theory exist. If there are successful theories of cognitive capacities which postulate representations of whose contents we are not aware then these representations are thought to exist.

He cites Marr's theory of vision (1982) as an instance of such a theory, and the image and primal sketch as examples of putative unconscious (explicit) representations. Schröder then asks: "Shall we try to do without [these representations] just because our favorite theory of consciousness says there cannot be such things?" **O'Rourke** (implicitly) raises the same issue when he challenges us to provide an explanation of delayed recall without invoking explicitly represented information as part of the unconscious search process. Likewise, implicit in **Mortensen's** commentary is the idea that Freudian psychology depends on a causally efficacious unconscious. An inference to the best explanation takes one quickly to the view that the unconscious is populated by a multitude of explicitly represented beliefs and desires.

We think **Schröder** is unduly dismissive of the obligations that a theory of consciousness might place on theories of perception or cognition. A theory of consciousness ought to *constrain* our theorizing in other areas, especially when it takes a computational form. The ultimate science of the mind will be an integrated package, achieved by a triangulation involving the neurosciences, computational theory, and first-person conscious experience. So if our theory of vision does not cohere with our best account of consciousness, it is back to the drawing board, and there is no *a priori* argument to the effect that it is the theory of consciousness that will have to go.

However, the central point here is well taken. Where non-contrastive analyses have been successfully applied to cognitive phenomena we surely have some reason to take seriously any unconscious, explicitly represented information they appeal to. That said, it is difficult to know how to assess this objection. Until recently, theorizing in cognitive science has been dominated by the classical conception of cognition (this is certainly true of Marr's theory of vision). We have argued that classicists are committed to the existence of explicit representations whose contents are not conscious (target article, Section 3.2), so it is no surprise that such representations are

legion in current theorizing. An inference to the best explanation is always vulnerable to the emergence of a rival theory with comparable simplicity and explanatory scope. With the advent of connectionism a whole new *class* of explanations is coming onto the scene, and it is no longer safe to assume that classically inspired theories will retain their favored status.

More importantly, from the perspective of a vehicle theorist, it is no longer safe to assume that theories in cognitive science will continue to rely on a classical-style unconscious, given the role of nonexplicit information in PDP systems. In particular, it is not clear that connectionist models need invoke anything like the number of explicit representations employed in traditional models of perception. NETtalk is paradigmatic in this regard (Sejnowski and Rosenberg, 1987). NETtalk takes English language text as input and produces its phonemic analysis, doing so with a high degree of accuracy and reliability. A conventional implementation of this task (such as Digital Equipment Corporation's DECtalk) requires hundreds of complex conditional rules, and long lists of exceptions. By contrast, NETtalk employs *no* explicit rules and *no* explicit data storage. It transforms input to output via a single-step process that is, in effect, an extremely complex form of (contextually nuanced) pattern-matching.

Of course NETtalk is a minuscule network, by brain standards. But when PDP techniques are applied to large-scale perceptual systems (such as the visual system[6]), the moral seems to be the same: whereas symbol-processing models invariably appeal to explicit rules and exception classes, PDP models invoke complex nets, or hierarchies of nets, that process input in a monolithic, reflex-like fashion. Such processing generates the contents of perceptual experience without any explicit unconscious intermediaries. A similar approach may be taken to cognition in general: the "psychological reflexes" that enable a chess player to "see" the best move, and a native speaker to effortlessly parse speech signals; and the extended bouts of conscious thought characteristic of calculation, reasoning, and creative thinking. The vehicle theorist pictures these as monolithic relaxation processes, or as a hierarchically-chained sequence of such computations (in the case of extended reasoning), which harness the brain's vast store of nonexplicit information.[7]

Obviously this is all very promissory. The onus is clearly on a vehicle theorist to provide PDP models of this type. We have begun this task elsewhere (Opie 1998), but a great deal remains to be done.

6. The Dissociation Studies Revisited

Unsurprisingly, a number of commentators have taken us to task over our treatment of the dissociation studies: that large body of work which is widely interpreted as having established the existence of unconscious, explicitly represented information. **Perner & Dienes** claim that our review of the literature is "selective and dated", a sentiment echoed by **Velmans**. These commentators argue that the evidence for dissociation between conscious experience and explicit mental representation is more compelling than we allow, and point to specific studies we neglected. Velmans makes the additional point that we are burdened with demonstrating that *all* of the contrastive analyses conducted to date are flawed, thus "even one good example of preconscious or unconscious semantic processing would be troublesome for [the connectionist vehicle theory of consciousness]".

---

[6] See, for example, Arbib & Hanson 1987, and Lehky & Sejnowski 1990.

[7] This is probably not the way to tackle delayed recall. One promising approach is suggested by the work of Smith and Blankenship (1989, 1991). Their examination of unconscious incubation, which is of a piece with delayed recall, suggests that incubation amounts to no more than the gradual decay of a memory block. This explanation is essentially non-cognitive, in that it doesn't appeal to processes defined over information bearing states. For other explanations of incubation along these lines see Boden 1990, pp.244-5, and Perkins 1981.

These objections are well taken. However, in our defense, we took our role in the target article to be one of establishing that a connectionist vehicle theory of consciousness is not *conclusively* ruled out by existing studies. In this connection, it is interesting to note the rather different appraisal of the literature offered by **Dulany**, who describes the dissociation studies as being subject to "decades of conceptual confusion and methodological bias". He continues:

> Suffice it to say now that if claims for the power of a cognitive unconscious were correct, the experimental effects would be too strong and replicable for these literatures even to be controversial. No one can claim that.

This level of disagreement among cognitive psychologists certainly suggests that it is legitimate to pursue a vehicle theory, while suspending judgement regarding the dissociation studies. Nevertheless, **Velmans** is right about the burden we must ultimately shoulder. The onus is on a vehicle theorist to show that each of the contrastive paradigms is flawed in some way, or is open to reinterpretation in the light of connectionism. So we will make some further remarks about implicit learning, take a second look at the phenomenon of priming, and finish with a discussion of the more recent literature on blindsight.

6.1 Implicit Learning

In our discussion of implicit learning we relied heavily on the work of Shanks and St. John (1994), who characterize this phenomenon as the induction of unconscious rules from a set of rule-governed training stimuli. Perner and Dienes's commentary suggests, and a survey of the more recent literature confirms, that this is not the only way theorists are apt to characterize implicit learning (see, e.g., Cleeremans 1997, Dienes & Berry 1997, Perruchet & Gallego 1997, Perruchet & Vinter 1998). A less contentious way of defining implicit learning, inspired by Perruchet & Gallego 1997 (p.124), is:

> an adaptive process whereby subjects become sensitive to the structural features of some stimulus domain without consciously deploying learning strategies to do so.[8]

This definition captures what occurs in the standard experimental paradigms designed to investigate implicit learning: artificial grammar learning, instrumental learning, serial reaction time learning, and so on. However, it is sufficiently generic to cover aspects of first and second language learning, the acquisition of reading and writing, and adaptation to physical and social constraints. The essential contrast is with cases where a subject is aware that learning is taking place, and deploys various strategies to facilitate the process (for example, consciously forming and testing hypotheses about the stimulus domain).

In light of this definition the acquisition of abstract, unconscious rules is best seen as one among a number of possible *explanations* of implicit learning. This approach—best exemplified in the work of Reber (1993) and Lewicki (1986)—is very much classically inspired, as Cleeremans points out. It is clearly incompatible with the connectionist vehicle theory of phenomenal experience, since it assumes the operation of explicitly represented information that does not figure in consciousness.

However, denying a particular explanation of implicit learning does not amount to denying the existence of the phenomenon. The issue for us now becomes: Is there an explanation of implicit learning (as defined above) that is compatible with the connectionist vehicle theory of consciousness? We think the answer is yes. The shape of such an explanation is suggested by Perner & Dienes, and finds its most detailed elaboration (minus the connectionist gloss) in the work of Perruchet and colleagues (Perruchet & Gallego 1997, Perruchet & Vinter 1998). They propose a *subjective unit-formation* account of implicit learning, in which "intrinsically unconscious" associative mechanisms generate increasingly appropriate parsings of the stimuli

[8] See also Cleeremans 1997, Section 3.

in some domain. First exposure to stimulus material brings on line specialized mechanisms—which may be innate, or the product of earlier learning—that parse stimuli into a set of small disjunctive units. These subjective units comprise our experience of the domain. They are selected and modified by subsequent training, and can go on to form the basis of higher-level subjective units (which fits with the way training both focuses and enriches our experience). (See Perruchet & Vinter 1998, pp.502-5 for a summary of this account.)

We believe the connectionist vehicle theory of consciousness complements this account of implicit learning. The subjective units produced at each stage, being conscious, may be envisaged as stable activation patterns. Continued exposure to the stimulus domain will therefore initiate the learning mechanisms proposed by Perruchet and colleagues, be they Hebbian or otherwise, since stable signalling among networks is surely vital to the modification of connection weights. Such modifications, in their turn, will alter the subjective units, because connection weights control the relaxation processes that generate conscious experience. We can thus understand how it is that experience shapes learning, and learning in its turn alters experience.[9] There is the appearance of a disagreement between us, because **Perruchet & Vinter**, in their commentary, take us to be identifying consciousness with the result of learning: a network whose *connection weights* have stabilized. In other words, they interpret stability *diachronically*. But, in actual fact, we are offering a *synchronic* hypothesis: conscious experiences are identical with *stable activation* in networks that *have already been trained*. This misunderstanding is probably partly due to our failure to say enough about learning.

6.2 Subliminal Priming

Several commentators raise concerns about our treatment of subliminal priming. **Perner & Dienes** claim that our critique of Marcel's (1983) experiments "rehashes old arguments already dealt with [by Marcel]", and that we have failed to mention more recent studies. **Velmans** draws attention to the work of Groeger (1984) who found evidence of semantic priming, on a subsequent word selection test, by subliminally presented words. He also refers to studies which appear to demonstrate that attention to a spoken word preconsciously activates all of its possible meanings for a short period (around 250 milliseconds): "Depending on context, one meaning is selected and subsequent entry of the word into consciousness is accompanied by inhibition (or deactivation) of inappropriate meanings…" **Zorzi & Umilta**, in their thoughtful commentary, remind us of the priming studies that have been done with subjects suffering from unilateral neglect (Berti & Rizzolatti 1992, Làdavas et al 1993). For example, objects presented to the neglected hemifield of subjects with severe neglect appear to facilitate (i.e., speed up) responses to semantically related objects presented to the contralateral field.

In our target article, we show that it is reasonable to have doubts about the dissociation studies. Following Holender (1986), we therefore raise some methodological worries for visual masking, and, by implication, for other paradigms that investigate subliminal priming. But there is a stronger response we can make. Here we propose to take the various priming phenomena at face value, but then show how they may be explained in a way that is consistent with the connectionist vehicle theory of consciousness. In so doing we will be able to further illustrate a significant feature of information processing in PDP systems.

The reasoning behind the various priming studies seems to be that, in order for a subliminal stimulus to affect ongoing cognitive processing, an explicit representation of some sort has to be generated and manipulated in the cognitive system. If so, subliminal priming provides clear evidence for the existence of explicit representations with unconscious contents. But there is some question as to whether subliminal priming, *when interpreted from within the connectionist*

---

[9] In addition, Perruchet et al emphasize in their account the importance of hierarchical processing (see, for example, Perruchet & Vinter 1998, p.503), which we also take to be explanatorily significant (target article, Section 5.2).

camp, unequivocally leads to this conclusion. In particular, one should always be mindful of the *style* of computation employed by connectionist systems. A recurrent network, for example, when initially exposed to an input can oscillate quite dramatically as activation circulates around the network, and hence can take some time to relax into a stable pattern of activation (though here, of course, we are talking in terms of milliseconds). However, just prior to stabilization, as these oscillations abate, the network is likely to converge on some small region of activation space. It is this feature of PDP processing that provides the leverage for a connectionist explanation of subliminal priming.

Consider cases of priming in normal subjects (we will turn to the case of unilateral neglect shortly). Some priming studies find better than chance performance in forced-choice judgment tasks (say, comparing two stimuli which may be semantically related), even though initial stimuli are presented 5-10 msec below the supraliminal threshold. Other studies find that subliminal primes can facilitate the recognition of subsequent (supraliminal) stimuli. In many of these studies, the primed stimulus occurs immediately after, or within a very short period of, the priming stimulus. The following type of explanation is available in such cases. Because of the short duration of the initial stimulus there is not enough time for a stable pattern of activation (and thus an explicit representation) to be generated in the relevant networks. Thus, when a second stimulus enters the system it will interfere with processing that has already begun, but not gone to completion. Moreover, if this second stimulus is related to the priming stimulus in some cognitively salient way, then the likely effect of this interference will be rapid stabilization. As Zorzi & Umilta remark, related items are represented in PDP systems by similar activation patterns, and so a relaxation process that is near completion will already be in a region of activation space suitable to the representation of the second stimulus. Consequently, the PDP system will relax more quickly when the second stimulus is related to the prime than when it is unrelated. Crucial to this account, from our perspective, is the fact that the initial stimulus is *never explicitly represented*, because network relaxation doesn't go to completion until after arrival of the second stimulus. However, the first stimulus still influences the course of processing (assuming it is sufficiently close to the supraliminal threshold), via the relaxation process that it sets in train. This explains important cases of subliminal priming without invoking explicit, unconscious primes.[10]

When it comes to neglect we only need to alter this story a little. Although it is the presence of a pattern mask (or of some other stimulus) that ensures that primes are subliminal in normal subjects, in the case of subjects with unilateral neglect it is the damaged condition of the brain which explains the failure of information to enter consciousness. We might conjecture that such damage interferes with the capacity of networks in the neglected hemifield to settle into stable activation patterns. Nevertheless, signals that enter here may still have some influence on processing in the intact hemifield. This conjecture aside, it is important to remember that unilateral neglect studies don't provide unequivocal support for subliminal priming. One should always be mindful that where brain-damage is involved, failures to report awareness may be caused by communication breakdowns in the brain, rather than genuine dissociations between phenomenal experience and explicit representation. The difficulty of distinguishing between these two possibilities is particularly acute in the case of neglect. It is for this reason that Berti and Rizzolatti, instead of concluding that their subjects "showed a phenomenon akin to blindsight" prefer "a more parsimonious interpretation, namely that our patients had a severe neglect and *behaved* as if they had hemianopia without really being hemianopic" (1992, p.348).

---

[10] This explanation may require that we allow for the *co-stabilization* of neural networks, that is, of *inter*-network activation-passing that results in a synchronized approach towards stability. But this is surely a very common feature of neural computation, given the ubiquity of feedback connections in the brain, especially for networks in functionally related local groups.

Finally, where the "preconscious" activation of word meanings is concerned, connectionism provides a way of thinking about this phenomenon that dispenses with unconscious, explicit representations. Recall that in PDP networks information storage and information processing rely on the same substrate of connection weights and connections. It is the very word meanings encoded in a PDP system that determine how a lexical stimulus will be processed. But such word meanings have their effects without becoming explicit, and this explains why most of them don't enter consciousness, on our account. When one of these meanings does become conscious, this is not because its rivals are "inhibited" or "deactivated". It is rather that the relaxation process which constitutes most PDP computation is only capable of rendering explicit *one* among the great many meanings that are potentially explicit in the system. The general point here is that cognition, from the connectionist perspective, is a good deal more holistic than classicism allows. A great deal of information processing takes place in a connectionist network prior to the production of an explicit mental representation. Such processing can produce facilitation effects, in the manner described above, without the involvement of explicitly represented information.[11]

6.3 Blindsight

A lot has happened in blindsight research of late, as **Perner & Dienes**, and **Kentridge**, rightly point out. The stray light hypothesis, offered by Campion, Latto & Smith (1983) as a possible explanation of blindsight, appears to have been taken on board and controlled for (see, e.g., Milner & Goodale 1995, pp.72-3, and Shallice 1997, p.258 for discussion). The spared cortex hypothesis also looks less plausible in light of recent work (Kentridge, Heywood & Weiskrantz 1997). The evidence is now fairly conclusive that a range of visually guided behaviors can occur without striate mediation, and, as Kentridge points out, that both subcortical and cortical structures are implicated in these behaviors (see Tovée 1996, pp.71-4 for a brief discussion).

Having acknowledged these developments, we still have some serious concerns about blindsight research. Many investigators continue to disregard reports of "feelings" in blindsight subjects, and there is a general lack of consensus concerning the status of phenomenal reports. As we urge in our target article, in order to adequately assess the evidence for the dissociation of explicit representation and phenomenal experience it is crucial that *any phenomenal reports whatsoever* be taken into consideration. In addition, there is a persistent problem in the methodology of blindsight research: the use of *post-trial* subject reports to explore the relationship between performance and awareness. In some cases as many as several hundred visual presentations occur before awareness is assessed! It is remarkable that it was not until very recently that systematic attempts were made to investigate the relationship between awareness and performance on a trial-by-trial basis, while allowing for various levels of visual awareness, including reports of feelings (Weiskrantz et al 1995, Kentridge et al 1997, Zeki & ffytche 1998—see Cowey 1996 for a brief overview of other recent work).

What of these recent studies? All of them have involved testing a single subject, GY, whose left visual cortex is badly damaged, such that he is clinically blind in his right hemifield. Weiskrantz et al (1995) tested GY on a motion discrimination task using a two-alternative forced-response procedure. He responded by key press, using one of two keys to indicate the direction of motion of a moving spot presented in his blind hemifield, and one of two keys to indicate, respectively, no awareness, or some awareness. Awareness was tested after every trial, in blocks of 50 or 100 trials. On those trials where he signalled no awareness GY was still required to guess a direction

---

[11] Having conceded this, it would not do to exaggerate the significance of subliminal priming. Such effects only arise when the initial stimulus duration is just marginally below the supraliminal threshold, and involve small facilitations (e.g. a 5% decrease in a reaction time, a slightly above chance judgement in a forced-choice task, etc.). One might think that the limited nature of these priming effects should discourage researchers from inferring too precipitously that explicitly represented, unconscious information must be involved.

of motion. Striking results were obtained. Although GY often reported visual experiences of some kind, in those instances where he reported no awareness whatever he achieved as high as 90% accuracy (across a block of trials) for direction of motion judgements.

In a variation on this paradigm Zeki & ffytche (1998) introduced a four-level scheme for reporting awareness, ranging through: 1 - no awareness, 2 - feeling that something was there, but guessed the direction, 3 - fairly confident of the direction, 4 - certain of the direction. Zeki & ffytche found that performance generally correlated with awareness as one might expect (i.e., above chance performance corresponded with a high percentage of aware trials within a block, chance levels of performance corresponded with a low percentage of aware trials), but also discovered blocks of trials in which, with *no* reports of awareness across the block, the levels of performance were well above chance (on the order or 70 or 80% correct responses).[12] Again, these are quite striking results.

Prima facie these studies present something of a problem for our account of consciousness. However, we feel that it is still possible to cast some doubt on these results, or to favorably reinterpret them. Kentridge raises the intriguing possibility (which he motivates with a discussion of cortical color blindness) that blindsight performance can be attributed to the presence of two distinct pathways in the visual system: the dorsal and ventral streams, both of which originate in V1, but terminate in different loci (the posterior parietal cortex, and inferotemporal cortex, respectively). Milner & Goodale (1993, 1995) have proposed that the ventral stream is specialized for visual learning and recognition, while the dorsal stream is devoted to the visual control of action. The dorsal stream may continue to process the visual signal, despite damage to V1, because it receives a number of subcortical projections. Thus, just as wavelength information is used to extract form, but not color, in cortical color blindness (Heywood, Cowey & Newcombe 1994), so the visual signal might be used to generate visuomotor representations in the dorsal stream, without any corresponding representations in the ventral stream, in blindsight. Crucially, such representations may well be associated with some kind of visuomotor phenomenology (what Kentridge calls "an awareness of action"). Since the studies discussed above only appear to test for something that GY would identify as visual, they don't eliminate the possibility that GY's performance is explicable in terms of action generating modules which produce some non-visual (visuomotor) experience.[13]

Marcel's (1993) study of GY adds weight to the conviction that the relationship between performance and experience in GY is still to be fully explored. GY's task was simply to respond to a spot of bright light in his blind hemifield. In one experiment he was asked to respond using three different modalities at once: blinking, button pressing, and verbal. There were frequent dissociations between the modalities; GY would, for example, report 'yes' with his eye, but not with his finger. In a second experiment the various modalities were tested separately. Eight blocks of 40 trials were run for each, in which GY was asked to " 'Respond as soon as you have an impression of whether a light came on…' " (1993, p.170). In a second condition (the "guessing" condition) a further two blocks of trials were run in which GY was instructed to "*guess* whether a light had come on, irrespective of any conscious sensation" (p.171). Marcel found that in the first two blocks of trials the success rates were similar across modalities,

---

[12] They also found that on some blocks of trials, although awareness levels were generally 2 or more, the performance was significantly poorer than expected.

[13] That there is more to say about GY's phenomenology is suggested by a remark reported in Zeki & ffytche (1998). At one point, when presented with a low contrast stimulus, GY "spontaneously remarked that the awareness score here should be 'minus one or minus two'…" (p.30). Zeki and ffytche take this to imply that there might be "degrees of unawareness" for GY. This seems a very odd conclusion, and the idea of 'degrees of unawareness' hardly seems coherent. The more natural conclusion, one might think, is that even when GY reports a 1 (corresponding to no awareness) on the four-level scale, he actually does have some kind of phenomenology – phenomenology compared to which a genuine 'no awareness' state looks like a 'minus one or two'.

although performance was marginally better for the blink response. However, there was a marked training effect for the blink and button press. Before training the percentages of hits (correct positive responses) for the blink and button presses were 67.5 and 62.5, respectively. After training this figure had jumped to 70 for the button press, and 85 for the blink response! Verbal response did not improve with training, and performance stayed around 60% hits (1993, p.172). Most strikingly, in the guessing condition Marcel found that GY was able to achieve on the order of 80% hits (87.5 for the blink response), *irrespective of response modality*. In other words, GY's performance using button press and verbal assent when merely guessing, approached the results achieved using blink response when attending to visual experience.

How does one interpret these results? To begin with, it is reasonable to suppose that the improvement in blink response performance is reflected in alterations to structures in the dorsal stream, where blink responses are known to be controlled (see Milner & Goodale 1993). What is significant is that performance across all response modalities, under the instruction to *guess*, is no better than the trained-up performance for blinking, under the instruction to respond to a visual experience. The most parsimonious reading of these results is that the very phenomenology driving the blink response is also responsible for performance in the guessing condition. That is, guessing is actually guided by some sort of phenomenology, but phenomenology that only has an indirect effect on button presses and verbal responses. One might think, in fact, that the phenomenology in this case is visuomotor, given the seemingly crucial role of the dorsal stream in these results. Vision remarks that "[p]erhaps…GY did not feel compelled to perceive the light on the guessing trials, but only to use whatever feelings he had available for answers (even if they were based on feelings acquired from premotor muscular preparation)" (1998, p.151).

An implication of all this, and one that Marcel explicitly draws, is that phenomenal consciousness is not nearly so unified as we usually imagine. In the report condition the blink response is generally more effective than the others, because it appears to be controlled by a dedicated part of the visual system in the dorsal stream. The other modalities are down-stream of the visual system, and, given the damaged condition of GY's striate cortex, are less likely to get consistent signals from there. Consequently we get the odd dissociations between the response modalities noted above. Such disunity is not apparent in normal subjects, because the lines of communication among distinct response systems are generally good (although see Marcel 1993 for a control study on normal subjects with degraded—low luminance contrast—stimuli). This is an issue we take up in the next section.

7. Subject Unity, Agency and Introspection

Your conscious experiences do not just occur, they occur *to you*. The multifarious perceptual and understanding experiences that come into being as you read these words are somehow stamped with your insignia. In the target article we call this subject unity, and note that given the multi-track nature of our vehicle theory—given that there are consciousness-making mechanisms scattered right across the brain—there is a real issue as to how subject unity arises (Section 5.3). We think this is one of the most difficult problems facing a general theory of consciousness, and a number of commentators have identified what they take to be inadequacies in our treatment. In particular, **Carlson**, although he believes we offer "an intriguing hypothesis" about the contents of phenomenal experience—the "what" of consciousness—thinks we fail to properly address subjectivity and conscious agency—the "who" of consciousness. Likewise, **Dulany** raises concerns about sense of agency, metacognitive awareness, and consciousness of self. And **Schwitzgebel** argues that neither the existence of a narrative, nor the confluence of points of view, can successfully explain our sense of subject unity. These issues are extremely important, and we accept that our treatment of them in the target article was cursory and incomplete. We feel, nonetheless, that the connectionist vehicle theory points us in the right direction.

Carlson would like a theory of consciousness which can account for the "existence and activity of conscious agents", and which shows "how consciousness contributes to the control of purposive activity". It is important to distinguish the different demands being made here. To account for the *activity* of conscious agents is not the same thing as accounting for the *experience* of conscious agents, in particular the sense of self and sense of agency that partly constitute the conscious agent. Connectionism suggests that the activity of an agent results from the collective and cooperative operation of a great many PDP networks, and therefore that control is highly distributed in the brain. This idea coheres with the work of Marcel (1993), of Milner & Goodale (1995) (see previous section), and with a great deal of data regarding motor and cognitive deficits. It contrasts with the view that there is some central kernel—an executive—that directs and controls our cognitive and motor behavior.

A consequence of the distribution of control, if we take it seriously, is that agency must be seen as an emergent, perhaps having no locus smaller than the entire brain. The connectionist vehicle theory of consciousness suggests that our experience of agency (our experience of ourselves *as* agents) likewise emerges from the activity of a multitude of neural networks; that it is a sum of numerous distinct stable activation patterns. Indeed, assuming a vehicle theory, the distributed neural basis of consciousness is intimately related to the distributed nature of the agent (understood as the locus of control). This is because stable activation has such a crucial role in the inter-network communication that mediates control, enabling coherent activity to emerge from disparate sources. Therefore, although it is important to clearly distinguish the subject as actor (the active self) from the subject as experiencer (the phenomenal self), there is actually a very tight coupling between the two on our account. Contrary to what **Dulany** suggests, consciousness is very much driving the bus, because conscious states (stable activation patterns) are so bound up with the inter-network processing at the heart of both cognition and action.

This framework for issues surrounding self and agency immediately raises further questions and problems, the most significant of which are:

1) How, in detail, are the control elements coordinated?
2) What are the phenomenal elements that go to make up our sense of agency?

Regarding the latter, a preliminary analysis suggests that our sense of agency includes our sense of self, our awareness of our actions, and our awareness of the relationship between the two. Our sense of self, in turn, includes our bodily awareness, and our conscious plans and goals. Some will object that these phenomenal elements are far too abstract and ill-defined to map neatly onto patterns of activation in distinct neural networks. We accept this, and our initial characterisation clearly needs a great deal of further elaboration. **Mac Aogáin** appears to deny that such abstract elements feature in consciousness at all; the phenomenal world, as we describe it, is "too loosely defined to give decisive results". We contend, however, that phenomenology just does comprise a great many disparate elements. There is really no more problem accepting understanding experience as a genuine part of our phenomenology, than taking visual and auditory experiences to have something in common—they are clearly very different, yet we standardly treat them as members of a kind (see Flanagan 1992, Chp.4 for more on this theme).

As for the coordination of control, a proper answer to this question would essentially involve us in providing a complete connectionist account of cognition. Such an account would show, in detail, how perceptual, cognitive and motor processes are integrated, and would at every point indicate the role of phenomenal experience in this integration. It would also address many of the issues that rightly concern **Dulany**: the nature of propositional contents, deliberative thought, metacognitive awareness, and so on. We have done some preliminary work on these issues (Opie 1998, Chapt.7), but clearly much remains to be done.

McDermott expresses concern that on our account the link between introspection and consciousness is not necessary; that we allow for "the bizarre possibility that most of our

conscious experiences aren't accessible to introspection". If we follow McDermott, and treat introspection as the process whereby one conscious state becomes the object of another, then introspection is surely a common feature of human thought. However, this is no surprise on the connectionist vehicle theory of consciousness. Given the massive connectivity of the brain, and the role of stable activation patterns in inter-network processing, one would expect almost every phenomenal experience to be available, in principle, to introspection and verbal report. It is only under pathological, or degraded input conditions, that one would anticiapte any deviation from this.

McDermott's objection seems to be partly motivated by the sorts of intuitions that impress Higher-Order-Thought (HOT) theorists like **Perner & Dienes**. They think of consciousness as a mechanism that gives us *access* to our mental states, and so propose the following necessary condition for consciousness: X is conscious if there exists a *second order* state that represents the mental state with content X. Again, it is certainly true that human experience incorporates a great many higher-order phenomenal states. I can have a conscious perceptual experience, and I can simultaneously reflect on that experience (noting, for example, how *intense* the colors are, how *beautiful* the music is, and so forth). Here a number of networks are involved, including those that generate linguistic consciousness. But such higher-order thoughts are experiences too, so the distinction that inspires HOT theorists is a distinction *within* consciousness. For this reasons it strikes us as a singularly inappropriate basis for a story about the *constitution* of consciousness.

Similar confusions abound in the literature concerning the relationship between consciousness and attention. Some theorists conflate the two, and so take a theory of attention to be none other than a theory of phenomenal consciousness. Elsewhere we argue that this conflation is fundamentally mistaken (O'Brien & Opie 1998). Consciousness incorporates both a central focus, and a rich polymodal periphery. Theorists often neglect the periphery, but it is important for all that (it can save your life when crossing the road). Both focus and periphery are parts of one's instantaneous experience, so the attended/unattended distinction does not line up with the conscious/unconscious distinction, **Coltheart's** suggestions notwithstanding. The standard view, as we understand it, is that cognition depends on an enormous number of mental representations whose contents are not merely peripheral, but completely absent from consciousness. Incidentally, we can't agree with **Coltheart** that it only possible to attend to one thing at a time. One can, for example, hold a conversation and deal cards at the same time. Support for this intuition comes from recent studies which have led to the proposal of a multiple resource theory of attention (see Anderson 1995, pp.103-4 for discussion).

Finally, we turn to the difficult issue raised by **Schwitzgebel**. If the elements of consciousness are generated at multiple sites throughout the brain, as we contend, what is it that unifies these elements, such that they are all part of a *single* consciousness? Some experiences have a "togetherness" (such as my seeing your face and my hearing your voice) that others lack (such as my seeing your face and your hearing your own voice) (this example is adapted from Hurley 1993, p.50). Glover makes the same point when he asks us to consider:

> a procession, where half the people taking part are deaf and the other half are blind…There will be many visual experiences and at the same time many auditory ones. But none of this generates the unified experience of both seeing and hearing the procession. (1988, pp.54-5)

Having recognized this problem, it is still important to ask: What exactly needs explaining here? There seem to be two possibilities.

> 1) We need to account for our *sense* of unity—our sense of being a single, coherent, embodied self; of being the focus of action; of being an agent.

If this is the problem then our response to **Dulany** and **Carlson** is the beginning of a solution—one that relies on inter-network processing to maintain the coherence of experience, and to

generate the multiple abstract contents that constitute a sense of unity. Our suggestions about narrative and point of view may be seen as contributions to the task of analysing this 'sense of unity' (in its diachronic and synchronic forms, respectively).

2) We need to account for unity as an "ontological" feature of phenomenal consciousness.

This is tougher, because it requires that we explain the "togetherness" of certain experiences without treating this unity as a further (abstract) element of consciousness. **Schwitzgebel's** suggestions are welcome here. Advocating a vehicle theory of the contents of consciousness, does not, in our view, preclude one from proposing a theory of the unity of consciousness in which specific causal or information relations (of the kind only found within a single brain) are responsible for the way phenomenal experiences hang together.

References (Additional to those in the target article)

Anderson, J.R. (1995) Cognitive Psychology and its Implications, 4th edition. Freeman.

Arbib, M. & Hanson, A. (1987) Vision, Brain, and Cooperative Computation. MIT Press.

Baars, B.J. (1994) A thoroughly empirical approach to consciousness. Psyche 1(6)

Berti, A. & Rizzolatti, G. (1992) Visual processing without awareness: Evidence from unilateral neglect. Journal of Cognitive Neuroscience. 4:345-51.

Blachowicz, J. (1997) Analog representation beyond mental imagery. Journal of Philosophy 94: 55-84.

Boden, M. (1990) The Creative Mind. Abacus.

Churchland, P.M. (1990) Knowing qualia: A reply to Jackson. In: A Neurocomputational Perspective. MIT Press.

Clapin, H. & O'Brien, G. J. (1998) A conversation about superposition and distributed representation. Noetica: Open Forum, 3(10) http://psy.uq.edu.au/CogPsych/Noetica/

Cleeremans, A. (1997) Principles for implicit learning. In: How Implicit is Implicit Learning?, ed. D.C.Berry. Oxford University Press.

Cowey, A. (1996) Visual awareness: Still at sea with seeing? Current Biology. 6(1):45-7.

Cummins, R. (1989) Meaning and Mental Representation. MIT Press.

Dienes, Z. & Berry, D. (1997) Implicit learning: Below the subjective threshold. Pyschonomic Bulletin and Review. 4:3-23.

Edelman, S. (forthcoming). Representation is the representation of similarities. Behavioral and Brain Sciences.

Files, C. (1996) Goodman's rejection of resemblance. British Journal of Aesthetics 36: 398-412.

Gardenfors, P. (1996) Mental representation, conceptual spaces and metaphors. Synthese 106: 21-47.

Glover, J. (1988) I: The Philosophy and Psychology of Personal Identity. The Penguin Press.

Groeger, J.A. (1984) Evidence of unconscious semantic processing from a forced error situation. British Journal of Psychology. 75:305-14.

Heywood, C.A., Cowey, A. & Newcombe, F. (1994) On the role of parvocellular (P) and magnocellular (M) pathways in cerebral achromatopsia. Brain. 117:245-54.

Hurley, S. (1993) Unity and objectivity. Proceedings of the British Academy. 83:49-77.

Jackson, F. (1982) Epiphenomenal qualia. Philosophical Quarterly 32: 127-36.

Kentridge, R.W., Heywood, C.A. & Weiskrantz, L. (1997) Residual vision in multiple retinal locations within a scotoma: Implications for blindsight. Journal of Cognitive Neuroscience. 9:191-202.

Kirsh, D. (1990) When is information explicitly represented? In: Information, Language, and Cognition, ed. P.Hanson. University of Bristish Columbia Press.

Làdavas, E., Paladini, R. & Cubelli, R. (1993) Implicit associative priming in a patient with left visual neglect. Neuropsychologia. 31:1307-20.

Lehky, S.R. & Sejnowski, T.J. (1990) Neural network model of visual cortex for determining surface curvature from images of shaded surfaces. Proceedings of the Royal Society of London B. 240:51-78.

Lewicki, P. (1986) Nonconscious Social Information Processing. Academic Press.

Lewis, D. (1990) What experience teaches. In: Mind and Cognition, ed. W. Lycan. Blackwell.

Marcel, A.J. (1993) Slippage in the unity of consciousness. In: Experimental and Theoretical Studies of Conciousness, Ciba Foundation Symposium 174, eds. G.R.Block & J.Marsh. John Wiley and Sons.

Marr, D. (1982) Vision: A Computational Investigation into the Human Representation and Processing of Visual Information. Freeman.

Milner, A.D. & Goodale, M.A. (1993) Visual pathways to perception and action. In: Progress in Brain Research, Vol.95, eds. T.P.Hicks, S.Molotchnikoff & T.Ono. Elsevier

Milner, A.D. & Goodale, M.A. (1995) The Visual Brain in Action. Oxford University Press.

Nemirow, L. (1990) Physicalism and the cognitive role of acquaintance. In: Mind and Cognition, ed. W. Lycan. Blackwell.

O'Brien, G.J. (forthcoming). Connectionism, analogicity and mental content. Acta Analytica.

O'Brien, G.J. & Opie, J. (1997) Cognitive science and phenomenal consciousness: A dilemma, and how to avoid it. Philosophical Psychology 10: 269-86.

O'Brien, G.J. & Opie, J. (1998) The disunity of consciousness. The Australasian Journal of Philosophy.

O'Brien, G.J. & Opie, J. (forthcoming) A defense of Cartesian materialism. Philosophy and Phenomenological Research.

Opie, J. (1998) Consciousness: A Connectionist Perspective. Doctoral thesis, University of Adelaide.

Perkins, D.N. (1981) The Mind's Best Work. Harvard University Press.

Perruchet, P. & Gallego, J. (1997) A subjective unit formation account of implicit learning. In: How Implicit is Implicit Learning?, ed. D.C.Berry. Oxford University Press.

Perruchet, P. & Vinter, A. (1998) Learning and development. In: Implicit Learning: Representation and Process, eds.P.Frensch & M.Stadler. Lawrence Erlbaum.

Reber, A.S. (1993) Implicit Learning and Tacit Knowledge. Oxford University Press.

Searle, J.R. (1992) The Rediscovery of Mind. MIT Press.

Shallice, T. (1997) Modularity and consciousness. In: The Nature of Consciousness, eds. N.Block, O.Flanagan & G.Güzeldere. MIT Press.

Shepard, R. and Chipman, S. (1970) Second-order isomorphism of internal representations: Shapes of states. Cognitive Psychology 1: 1-17.

Shepard, R. and Metzler, J. (1971) Mental rotation of three-dimensional objects. Science 171: 701-3.

Smith, S.M. & Blankenship, S.E. (1989) Incubation effects. Bulletin of the Psychonomic Society. 27:311-14.

Smith, S.M. & Blankenship, S.E. (1991) Incubation and the persistence of fixation in problem solving. American Journal Of Psychology. 104:61-87.

Tovee, M.J. (1996) An Introduction to the Visual System.

Vision, G. (1998) Blindsight and philosophy. Philosophical Psychology. 11:137-59.

Weiskrantz, L., Barbur, J.L. & Sahraie, A. (1995) Parameters affecting conscious versus unconscious visual discrimination with damage to the visual cortex (V1). Proceedings of the National Academy of Science of the United States of America. 92:6122-26.

Zekil, S. & ffytche, D.H. (1998) The Riddoch syndrome: Insights into the neurobiology of conscious vision. Brain. 121:25-45.