A Connectionist Theory of Phenomenal Experience

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**:127-48 (1999)

Abstract (Long)

When cognitive scientists apply computational theory to the problem of phenomenal consciousness, as many of them have been doing recently, there are two fundamentally distinct approaches available. Either consciousness is to be explained in terms of the nature of the representational vehicles the brain deploys; or it is to be explained in terms of the computational processes defined over these vehicles. We call versions of these two approaches *vehicle* and *process* theories of consciousness, respectively. However, while there may be space for vehicle theories of consciousness in cognitive science, they are relatively rare. This is because of the influence exerted, on the one hand, by a large body of research which purports to show that the explicit representation of information in the brain and conscious experience are *dissociable*, and on the other, by the *classical* computational theory of mind – the theory that takes human cognition to be a species of symbol manipulation. But two recent developments in cognitive science combine to suggest that a reappraisal of this situation is in order. First, a number of theorists have recently been highly critical of the experimental methodologies employed in the dissociation studies – so critical, in fact, it's no longer reasonable to assume that the dissociability of conscious experience and explicit representation has been adequately demonstrated. Second, classicism, as a theory of human cognition, is no longer as dominant in cognitive science as it once was. It now has a lively competitor in the form of *connectionism;* and connectionism, unlike classicism, does have the computational resources to support a robust vehicle theory of consciousness. In this paper we develop and defend this connectionist vehicle theory of consciousness. It takes the form of the following simple empirical hypothesis: *phenomenal experience consists in the explicit representation of information in neurally realized PDP networks*. This hypothesis leads us to re-assess some common wisdom about consciousness, but, we will argue, in fruitful and ultimately plausible ways.

Abstract (Short)

There are two fundamentally distinct computational approaches to phenomenal consciousness: either consciousness depends on the nature of the representational vehicles the brain deploys; or it is a product of special processes defined over these vehicles. We call versions of these two approaches *vehicle* and *process* theories, respectively. Process theories dominate the recent literature, but this orthodoxy is imposed on cognitive science largely by the *classical* computational theory of mind. *Connectionists*, on the other hand, are in a position to explore a vehicle theory of phenomenal experience. In this paper we develop and defend this vehicle theory. We show that while it leads us to re-assess some common wisdom about consciousness, it does so in fruitful and ultimately plausible ways.

1 Computational Theories of Consciousness: Vehicle versus Process

There is something it is like to be you. Right now, for example, there is something it is like for you to see the shapes, textures and colors of these words, to hear distant sounds filtering into the room where you sit, to feel the chair pressing against your body, and to understand what these sentences mean. In other words, to say that there is something it is like to be you is to say that you are phenomenally conscious: a locus of phenomenal experiences. You are not alone in this respect, of course, as the vast majority of human beings have such experiences. What's more, there's probably something it is like to be a dog, and perhaps even fish have phenomenal experiences, however minimal and fleeting these may be. On the other hand, there is surely absolutely nothing it is like to be a cappuccino, or a planet, or even an oak tree. These, at least, are the standard intuitions.[1]

It is clearly incumbent on any complete theory of the mind to explain phenomenal experience. And given that our best theory of the mind will likely issue from cognitive science, it seems incumbent on this discipline, in particular, to provide such an explanation. What is special about cognitive science is its commitment to the *computational theory of mind*: the theory that treats human cognitive processes as disciplined operations defined over neurally realized representations.[2] From this perspective, the brain is essentially a very sophisticated information processing device; or better, given what we know about brain architecture, an elaborate network of semi-independent information processing devices.

The computational vision of mind and cognition is by now very familiar. The question we want to consider here is how we might exploit the resources of this paradigm to explain the facts of phenomenal consciousness. Given that computation is information processing, and given that information must be *represented* in order to be processed, an obvious first suggestion is that phenomenal consciousness is somehow intimately connected with the brain's representation of information. The intuition here is that phenomenal experience typically involves consciousness "of something", and in being conscious of something we are privy to information, either about our bodies or the environment. Thus, perhaps phenomenal experience is the mechanism whereby the brain represents information processed in the course of cognition.

But to *identify* consciousness with the mental representation of information is to assert two things: that all phenomenal experience is representational; and that all the information encoded in the brain is phenomenally experienced. And theorists have difficulties with both aspects of this identification. On the one hand it is commonplace for philosophers to argue that certain kinds of phenomenal experience are not representational (John Searle, e.g., cites pains and undirected emotional experiences in this regard (1983, pp.1-2)); and on the other, it is sheer orthodoxy in cognitive science to hold that our brains represent far more information than we are capable of experiencing at any one moment in time. So sensations, undirected emotions and memories immediately pose problems for any account that baldly identifies phenomenal consciousness with mental representation.

---

[1] In speaking of 'phenomenal experiences' our intended target is neither self-consciousness nor what has come to be called access-consciousness (see Block 1993, 1995). It is, rather, phenomenal consciousness: the "what it is like" of experience (see Nagel 1974). We will speak variously of 'phenomenal experience', 'phenomenal consciousness', 'conscious experience', or sometimes just plain 'consciousness', but in each case we refer to the same thing.

[2] This description is deliberately *generic*. Some writers tend to construe the computational theory of mind as the claim that cognitive processes are the rule-governed manipulations of internal symbols. However, we will take this narrower definition to describe just one, admittedly very popular, species of computational theory, viz: the *classical* computational theory of mind. Our justification for this is the emerging consensus within cognitive science that computation is a broader concept than symbol manipulation. See, e.g., Cummins and Schwarz, 1991, p.64; Dietrich, 1989; Fodor, 1975, p.27; and Von Eckardt, 1993, pp.97-116.

The advocate of a such an account of consciousness is not completely without resources here, however. With regard to the first difficulty, for instance, there are some philosophers who, contrary to the traditional line, defend the position that all phenomenal experience is representational to some degree (we have in mind here the work of Tye (1992, 1996, forthcoming) and especially Dretske (1993, 1995)). The general claim is that the quality of our phenomenal experience, the what-it-is-likeness, is actually constituted by the properties that our bodies and the world are represented as possessing. In the case of pains and tickles, for example, it is possible to analyse these in terms of the information they carry about occurrences at certain bodily locations (see, e.g., Tye 1996). And as for the so-called "undirected" emotions, it is plausible to analyse these as complex states that incorporate a number of more basic representational elements, some of which are cognitive and some of which carry information about the somatic centres where the emotion is "felt" (see, e.g., Charland 1995; Johnson-Laird 1988, pp.372-376; and Schwartz 1990).

Moreover, with regard to the second difficulty, while it is undeniable that our brains unconsciously represent a huge amount of information, there is an obvious modification to the initial suggestion that might sidestep this problem. It is commonplace for theorists to distinguish between *explicit* and *inexplicit* forms of information coding. Representation is typically said to be explicit if each distinct item of information in a computational device is encoded by a physically discrete object. Information that is either stored dispositionally or embodied in a device's primitive computational operations, on the other hand, is said to be inexplicitly represented.[3] It is reasonable to conjecture that the brain employs these different styles of representation. Hence the obvious emendation to the original suggestion is that consciousness is identical to the *explicit* coding of information in the brain, rather than the representation of information *simpliciter*.

Let's call any theory that takes this conjecture seriously a *vehicle* theory of consciousness. Such a theory holds that our phenomenal experience is identical to the *vehicles of explicit representation* in the brain. An examination of the literature reveals, however, that vehicle theories of consciousness are exceedingly rare. Far more popular in cognitive science are theories that take phenomenal consciousness to emerge from the computational activities in which these representational vehicles engage.[4] These typically take the form of *executive* models of consciousness, according to which our conscious experience is the result of a superordinate computational process or system that privileges certain mental representations over others. Bernard Baars' "Global Workspace" model of consciousness (1988) is a representative example. Baars' approach begins with the premise that the brain contains a multitude of distributed, unconscious processors all operating in parallel, each highly specialized, and all competing for access to a global workspace – a kind of central information exchange for the interaction, coordination, and control of the specialists. Such coordination and control is partly a result of restrictions on access to the global workspace. At any one time only a limited number of specialists can broadcast global messages (via the workspace), since different messages may often be contradictory. Those contents are conscious whose representational vehicles gain access to the global workspace (perhaps as a result of a number of specialists forming a coalition and ousting their rivals) and are subsequently broadcast throughout the brain (pp.73-118). The nature of the vehicles here is secondary; what counts, so far as consciousness is concerned, is access to the global workspace. The emphasis here, is on what representational vehicles *do*, rather than what they *are*. The mere existence of an explicit representation is not sufficient for consciousness; what matters is that it perform some special computational role, or be subject to

---

[3] See, e.g., Dennett 1982; Pylyshyn 1984; and Cummins 1986. We discuss the distinction between explicit and inexplicit representation more fully in Section 3.

[4] See, e.g., Baars 1988; Churchland 1995; Crick 1984; Dennett 1991; Flanagan 1992; Jackendoff 1987; Johnson-Laird 1988; Newman 1995; Kinsbourne 1988, 1995; Mandler 1985; Rey 1992; Schacter 1989; Shallice 1988a, 1988b; and Umilta 1988.

specific kinds of computational processes. We shall call any theory that adopts this line a *process* theory of consciousness.

Why do process theories of consciousness dominate discussion in cognitive science? Or to put this round the other way: given that there are two quite different explanatory strategies available to cognitive scientists – one couched in terms of the representational vehicles the brain deploys, the other in terms of the computational processes defined over these vehicles[5] – why do so few chose to explore the former path?

The answer, we suggest, is twofold. First, there is the influence exerted by a large body of research which purports to show that the explicit representation of information in the brain and conscious experience are *dissociable*, in the sense that the former can and often does occur in the absence of the latter. We have in mind here experimental work employing such paradigms as dichotic listening, visual masking, and implicit learning, as well as the investigation of neurological disorders such as blindsight. Such "dissociation studies", as we'll call them, appear to rule out a vehicle theory. And second, there is the influence exerted in cognitive science by the *classical* computational theory of mind – the theory that takes human cognition to be a species of symbol manipulation. Quite apart from the dissociation studies, it has simply been a working assumption of classicism that there are a great many unconscious, explicit mental states. Indeed, we shall argue that classicism doesn't have the computational resources to defend a vehicle theory of consciousness – something that most theorists at least implicitly recognize. Thus, classicism and the dissociation studies form a perfect alliance. Together they have created a climate in cognitive science that inhibits the growth of vehicle theories. It is not surprising, therefore, that process theories of consciousness flourish in their stead.

But recent developments in cognitive science combine to suggest that a reappraisal of this situation is in order. On the one hand, a number of theorists have recently been highly critical of the experimental methodologies employed in the dissociation studies. So critical, in fact, that it's no longer reasonable to assume that the dissociability of conscious experience and explicit representation has been adequately demonstrated (see, e.g., Campion, Latto & Smith 1983; Dulany 1991; Holender 1986; and Shanks & St. John 1994.) And on the other, classicism, as a theory of human cognition, is no longer as dominant in cognitive science as it once was. As everyone knows it now has a lively competitor in the form of *connectionism*.[6] What is not so widely appreciated is that when we take a fresh look at these issues from the connectionist perspective, we find the terrain has changed quite considerably. Specifically, connectionism *does* have the computational resources to support a robust vehicle theory of consciousness, or so we shall argue.

Our primary aim in this paper is to develop and defend this connectionist vehicle theory of consciousness. We begin, in Section 2, with a rapid re-evaluation of the dissociation studies. It is not our goal here to provide a thorough-going refutation of this research, but, rather, to summarize some important criticisms that have recently been directed at it, and thereby undermine the view that the dissociation of consciousness and explicit representation has been

---

[5] Strictly speaking, there is a third alternative here, one that combines these two strategies. On this view, consciousness is to be explained in terms of both the intrinsic properties of the brain's explicit representational vehicles together with special kinds of computational processes defined over these vehicles. An application of the principle of parsimony suggests, however, that such a hybrid approach should be deferred at least until the other two explanatory strategies have been properly explored. Our concern is that while process theories have been much debated in cognitive science, vehicle theories have not yet been investigated in any real depth. We aim, in this paper, to raise the profile of this alternative strategy.

[6] We are assuming here that connectionism does constitute a *computational* account of human cognition (and is hence a competing paradigm *within* the discipline of cognitive science). Although some have questioned this assumption, we think it accords with the orthodox view (see, e.g., Cummins & Schwarz 1991; Fodor & Pylyshyn 1988; and Von Eckardt 1993, Chp.3).

conclusively established. This, we believe, provides some elbow room for exploring the possibility of a vehicle theory, a task we pursue in the remainder of the paper. In Sections 3 and 4 we examine the nature of information coding in classicism and connectionism, respectively, in an effort to determine whether either of these conceptions of cognition has the computational resources to support a vehicle theory of phenomenal consciousness. We conclude that such a theory is unavailable to classicists. But the same does not apply to connectionists. In the final substantive section of the paper (Section 5) we present and defend a connectionist vehicle theory of consciousness. This theory leads us to re-assess some common wisdom about consciousness, but, we will argue, in fruitful and ultimately plausible ways.

2 The Dissociation Studies: A Reappraisal

The literature in cognitive science is full of experimental work which claims to exhibit the dissociation of conscious experience and mental representation. The most influential paradigms are: *dichotic listening* and *visual masking*, which are reputed to provide good evidence for preconscious semantic processing; *implicit learning*, in which unconscious processes appear to generate unconscious rule structures; and studies of *blindsight*. This last, unlike the rest, is conducted with subjects who have damaged brains (specifically, ablations of striate cortex). All these paradigms are what Dulany calls "contrastive analyses", since they examine differential predictions concerning the existence and role of unconscious information in various kinds of thought (1991, p.107). And the almost unanimous conclusion derived from these studies is that human cognition implicates a great many representations that are both *explicit* and *unconscious*. In what follows we present a brief survey of this experimental work, with a view to raising some doubts about its methodological credentials.

2.1 Dichotic Listening

In dichotic listening tests subjects are simultaneously presented with two channels of auditory input, one per ear, and asked to perform various tasks. Early work within this paradigm was designed to study the nature and limits of attention (Baars 1988, pp.34-5). It was soon discovered, however, that information in an unattended channel can have effects on behavior. Results like these stimulated further research specifically aimed at investigating perceptual processes that occur without accompanying conscious awareness. This research falls into two major subgroups: *disambiguation* studies and *electrodermal response* studies. We won't consider the latter here, but see Holender (1986) for discussion and critique.

Lackner and Garrett (1972), and MacKay (1973) have done influential work based on the potential for disambiguation of information presented in the primary (attended) channel by information presented in the secondary (unattended) channel. Lackner and Garrett asked their subjects in a dichotic listening test to attend solely to the verbal input in the primary channel and paraphrase the sentences as they were presented. These sentences contained different kinds of ambiguities (i.e., lexical, surface structural and deep structural), and as they were presented a concurrent disambiguating context was presented in the secondary channel. Lackner and Garrett found that: "The bias contexts exerted a strong influence on the interpretation of all ambiguity types" (1972, p.365). Post-experimental subject reports indicated that "none of the subjects had noticed that the material being paraphrased was ambiguous" and "none of the subjects could report anything systematic about the material in the unattended ear" (1972, p.367). MacKay used a similar procedure, but instructed the experimental subjects to shadow the input to the primary channel (i.e., repeat it, word for word, while listening). One or two disambiguating words were presented in the secondary channel simultaneously with the ambiguous portion of the sentences in the primary channel, but apart from this the secondary channel was silent. MacKay also observed a strong bias towards the interpretation suggested by the disambiguating context (reported in Holender 1986).

The moral here is fairly obvious. In order to bias a subject's paraphrase of attended material, the unattended input must clearly undergo processing all the way to the semantic level. And if the unattended input is subject to this degree of processing it is reasonable to suppose that it has generated explicit mental representations somewhere in the brain. Yet both the Lackner and Garrett, and the MacKay studies suggest that this representation does not evoke any conscious experience. Thus, there is *prima facie* evidence for the dissociation of explicit representation and conscious experience.

However, not all cognitive psychologists accept the conclusions typically drawn from dichotic listening studies (see, e.g., Holender 1986). Indeed there is reason to believe that the apparent support for the dissociation generated by this research is an artefact of poor methodology. For example, there is the reliance on post-experiment verbal reports as a source of evidence for subjects' states of awareness during the trials. Nelson (1978) has demonstrated that verbal reports do not provide an exhaustive indicator of conscious awareness, because other tests, such as recognition tests, can detect items not revealed in verbal recall tests, while the converse is not true (reported in Shanks & St. John 1994). Equally problematic is the lack of control in relation to the allocation of attention. In the Lackner and Garrett studies there was no measure of subjects' actual deployment of attention, and Holender's analysis of the experimental protocols suggests that attention could not in fact have been fixed on the primary channel (1986, p.7). While MacKay's use of shadowing did provide a better control of the allocation of attention, it is known that attention can be attracted by isolated physical events in the secondary channel (Mowbray 1964). Most strikingly, in experiments designed to replicate the disambiguation effects, but in which attention deployment was better controlled, such effects did not appear (Johnston & Dark 1982; Johnston & Wilson 1980; Newstead & Dennis 1979). Thus, it is reasonable to conclude that the results obtained by Lackner and Garrett, and by MacKay, were entirely due to uncontrolled attention shifts to the secondary channel, shifts that resulted in brief conscious awareness of the disambiguating context, even if this experience couldn't later be recalled.

In response to this kind of criticism, Richard Corteen, one of the first theorists to develop and champion the dichotic listening paradigm (in electrodermal response studies), has issued the following reappraisal:

> I am convinced that the subjects in the Corteen and Wood (1972) study did not remember much about the irrelevant channel after the procedure was completed, but I have never been sure that they did not have some momentary awareness of the critical stimuli at the time of presentation…There seems to be no question that the dichotic listening paradigm is ill-suited to the study of unconscious processing, no matter how promising it may have appeared in the early 1970's (Corteen 1986, p.28, emphasis added).

2.2 Blindsight Studies

Among philosophers probably the best known experimental evidence for the dissociation of explicit representation and consciousness comes from "blindsight" studies. Weiskrantz coined this term to refer to visually guided behavior that results from stimuli falling within a scotoma (a blind part of the visual field) caused by ablations of striate cortex. (For a detailed examination of the phenomenon of blindsight, including both the historical background and more recent experimental developments, see Weiskrantz 1986). A number of studies indicate that subjects with striate ablations can localize flashes of light, or other visual objects, falling within a scotoma, which they indicate by pointing or by verbal distance estimate (e.g., Weiskrantz, Warrington, Sanders & Marshall 1974; Perenin & Jeannerod 1975, 1978; Weiskrantz 1980). There is also evidence that such subjects can discriminate patterns of various kinds. A forced-choice technique has been employed, in which subjects are presented with a succession of stimuli of varying orientations or shapes, and they must choose a pattern (from a range of possibilities provided to them) even when they claim not to see the object. Although the results here are quite

varied, with many subjects performing only at chance levels, Perenin (1978) found that some subjects could perform above chance, and Weiskrantz et al. (1974), using three pairs of stimuli, found that each of these two-way discriminations could be achieved, provided the stimuli were large, bright and of sufficient duration. (See Campion, Latto & Smith 1983 for a review of this literature.)

A principal claim of blindsight research is that it provides evidence for a subcortical system capable of giving rise to visually guided behavior. What has generated all the excitement among philosophers, however, is the further contention that such behavior can occur in the complete absence of visual phenomenology. Blindsight subjects frequently claim that they can't see anything, and that their answers in the forced-choice discrimination tests are merely guesses. It is this lack of visual awareness that presumably led Weiskrantz et al. (1974) to coin the term "blindsight". And it is this aspect of blindsight research that provides evidence for the dissociation of phenomenal experience and explicit representation. For it is reasonable to suppose that visual judgements are mediated by mental representations: in order for *anyone* to make discriminations concerning the visual environment, some sort of representation of that environment must first be generated. On the further assumption that such representations must be explicit (given that they are occurrent, causally active states), it appears that the phenomenon of blindsight constitutes evidence for dissociation.

However, one should not be too hasty here; blindsight research is not without controversy. Campion, Latto and Smith (1983) argue that none of the existing blindsight studies provides adequate controls for light scatter. Furthermore, they claim that it's impossible, on purely behavioral grounds, to distinguish between blindsight and vision mediated by degraded striate cortex, given the inherent unreliability of post-trial experiential reports (more on this shortly). Rather, "the issue of striate versus extrastriate mediation of function can only be satisfactorily solved, as in animal studies, by histological examination of the brain tissue" (p.445). In other words, studies to date haven't ruled out the following, more parsimonious hypothesis: that blindsight phenomena are the result of "light scatter into unimpaired parts of the visual field or…residual vision resulting from spared striate cortex" (p.423). Campion et al. support these claims with a number of experimental studies, in which they demonstrate the covariation of localization, awareness, and degree of light scatter in a hemianopic subject. Together with the methodological concerns raised above, and the failure to observe blindsight in cases of complete cortical blindness (p.445), these results suggest that a reappraisal of the orthodox interpretation of blindsight studies is in order.

There is thus reason to believe that blindsight depends, in one way or another, on processes mediated by striate cortex. Given that such processes normally lead to visual experience, this is somewhat puzzling, since blindsight subjects putatively have no visual experience of the objects they can localize, and/or identify. However, a solution to this puzzle is not hard to find, because it is with regard to this very issue that blindsight research is most seriously flawed. According to Campion et al. "there is wide disagreement about whether the subject is aware of anything at all, what he is aware of, and whether this is relevant to blindsight or not" (1983, p.435). Many authors assert that their subjects were not aware of any stimuli; others report various kinds and degrees of awareness; and some claim that nothing was "seen", but qualify this by conceding that their subjects occasionally do report simple visual sensations (pp.435-6). The disagreement here is probably partly due to equivocation over the use of terms like "aware" and "conscious" (among the researchers), in conjunction with a failure to ask precise enough questions of the experimental subjects. Weiskrantz acknowledges this difficulty: subject E.Y., when asked to report what he "saw" in the deficient half of his visual field, "was densely blind by this criterion", but "[if] he was asked to report merely when he was "aware" of something coming into his field, *the fields were practically full*" (Weiskrantz 1980, p.378, emphasis added).

When it comes to the substantive issue, it is essential that there be no equivocation: any reports of visual phenomenology, no matter how transient or ill-defined, seriously undermine the significance of blindsight for establishing dissociation. But in fact the literature contains a great many reports of experiences that co-occur with discriminative episodes. Consider the comments made by Weiskrantz' subject D.B., after performing well above chance in a test that involved distinguishing between Xs and Os presented in his scotoma. While D.B. maintained that he performed the task merely by guessing:

> If pressed, he might say that he perhaps had a "feeling" that the stimulus was either pointing this or that way, or was "smooth" (the O) or "jagged" (the X). On one occasion in which "blanks" were randomly inserted in a series of stimuli...he afterwards spontaneously commented he had a feeling that maybe there was no stimulus present on some trials. But always he was at a loss for words to describe any conscious perception, and repeatedly stressed that he saw nothing at all in the sense of "seeing", and that he was merely guessing (Weiskrantz et al. 1974, p.721).

Throughout D.B.'s verbal commentaries there are similar remarks. Although he steadfastly denies "seeing" in the usual way when presented with visual stimuli, he frequently describes some kind of concurrent awareness.

Consequently, while blindsight subjects clearly do not have normal visual experience in the "blind" regions of their visual fields, this is *not* to say that they don't have any phenomenal experience whatsoever associated with stimuli presented in these regions. What is more, it is not unreasonable to suggest that what little experience they do have in this regard explains their residual discriminative abilities. D.B., for example, does not *see* Xs or Os (in the conventional sense). But in order to perform this task he doesn't need to. All he requires is some way of discriminating between the two stimulus conditions – some broad *phenomenal* criterion to distinguish "Xness" from "Oness". And as we've seen, he does possess such a criterion: one stimulus condition feels "jagged" while the other feels "smooth". Thus, it is natural to suppose that he is able to perform as well as he does (above chance) *because of* the (limited) amount of information that is consciously available to him. We conclude that blindsight studies do not constitute good evidence for the extrastriate mediation of visual functions, and, more importantly, they do not provide any clear-cut support for the dissociation of conscious experience and explicit representation.

2.3 Implicit Learning

A further, very extensive literature that has an important bearing on the issue of dissociation concerns the phenomenon of implicit learning (see Dulany 1996, and Shanks & St. John 1994 for reviews). According to the standard interpretation, implicit learning occurs when rules are unconsciously induced from a set of training stimuli. This is to be contrasted both with conscious episodes of hypothesis formation and confirmation, and with memorizing instances (either consciously or unconsciously). A number of kinds of implicit learning have been investigated, including *instrumental learning*, *serial reaction time learning*, and *artificial grammar learning* (Shanks & St. John 1994). These studies all differ from those discussed above in that they concern relatively long-term alterations to reactive dispositions, as opposed to the short-term facilitations sought after in the dichotic listening and blindsight paradigms.

For our purposes it is obviously the claim that implicit learning is unconscious that is most significant, but some care needs to be taken in spelling out this claim. Most research on implicit learning has in fact been restricted to situations in which the training set is supraliminal (i.e., the stimulus durations and intensities are well in excess of those required to generate some phenomenology).[7] So it is not typically the *stimuli* that subjects are held to be unaware of in

---

[7] There has been some research on long-term priming in anaesthetized subjects, i.e., research involving subliminal stimuli, but this work is inconclusive (Shanks & St. John 1994, p.371).

implicit learning situations. It is, rather, the *relationships between the stimuli* that are thought to be unconscious (1994, p.371).

For example, consider the work on artificial grammar learning first conducted by Reber (1967). A typical experiment involves supraliminal exposure to a set of letter strings generated by a regular grammar (or, equivalently, a set of strings accepted by a finite automaton[8]), which subjects are asked to memorize, followed by a further set of novel strings which they must identify as either grammatical or ungrammatical. Subjects are generally able to perform well above chance on the grammaticality task, yet are unable to report the rules of the grammar involved, or indeed give much account of their decision-making. The standard interpretation of this result is that during training subjects unconsciously induce and store a set of rules. These rules are brought to bear in the grammaticality task, but do not enter consciousness (or, at least, are not reportable). There is *prima facie* evidence here that subjects exposed to training stimuli unconsciously acquire explicit knowledge of the relationships among those stimuli, which information guides subsequent decision-making, even though it remains unconscious.

It may that the standard interpretation is somewhat incautious, however. Shanks and St. John, in their wide-ranging critique, have identified two principal criteria which implicit learning studies must satisfy in order to establish unconscious learning (in the sense specified above). First, tests of awareness must be sensitive to all relevant conscious knowledge (the *sensitivity criterion*); and second, it must be possible to establish that the information the experimenter is seeking in awareness tests is actually the information responsible for changes in the subjects' performance (the *information criterion*). We won't consider the sensitivity criterion in detail here, but just note that a great many studies of implicit learning have relied entirely on post-experiment verbal reports, and this method of assessing awareness is known to be less sensitive than, for example, subject protocols generated during training, or recognition tests (see Shanks & St. John 1994, pp.374-5 for discussion). At any rate, it is the information criterion which appears to have been most deficient among those implicit learning studies that support the dissociability of phenomenal experience and explicit representation. When these studies are replicated it is repeatedly discovered that subjects do have some awareness of the relationships between stimuli.

In the artificial grammar learning studies, for example, Dulany, Carlson, and Dewey (1984) found that after learning "subjects not only classified strings by underlining the grammatical and crossing out the ungrammatical, but they did so by simultaneously marking features in the strings that suggested to them that classification"; moreover, subjects "reported rules in awareness, rules in which a grammatical classification is predicated of features" (reported in Dulany 1996, p.193). Similar results have been reported by Perruchet and Pacteau (1990), and Dienes, Broadbent and Berry (1991). In all of these studies subjects report the use of substring information to assess grammaticality (i.e., they recall significant pairs or triples from the training set, which they then look for in novel strings). Thus, a study that looks only for complex rules, or rules based on whole strings, will probably fail to report the kinds of awareness actually relevant to decisions regarding grammaticality; it will fail the information criterion.

Of particular significance is the finding that when reported rules are arrayed on a validity metric (which quantifies the degree to which these rules, if acted on, would yield a correct classification) they predict actual judgements "without significant residual"; even though "each rule was of limited scope, and most imperfect validity…in aggregate they were adequate to explain the imperfect levels of judgement found" (Dulany 1996, pp.193-4). Based on their extensive analysis of this literature, Shanks and St. John conclude:

---

[8] See Hopcroft & Ullman 1979 for the distinction between regular grammars (which Shanks and St. John call finite-state grammars) and finite automata. Regular grammars consist of a set of productions of the form $A \rightarrow w\ B$ or $A \rightarrow w$, where A and B are variables and w is a (possibly empty) string of symbols.

> These studies indicate that relatively simple information is to a large extent sufficient to account for subjects' behavior in artificial grammar learning tasks. *In addition, and most important, this knowledge appears to be reportable by subjects*. (1994, p.381, emphasis added)

They reach a similar verdict with regard to instrumental learning and serial reaction time learning (p.383, pp.388-9). It seems doubtful, then, that implicit learning, in the sense of unconscious rule-induction, has been adequately demonstrated at this stage. Just as in the case of blindsight, it appears that the (less than perfect) performance subjects exhibit in implicit learning tasks, can be fully accounted for in terms of information that is consciously available to them.

2.4 Visual Masking

Visual masking is one among a number of experimental paradigms employed to investigate subliminal perception: perceptual integrations that, due to short stimulus duration, occur below the threshold of consciousness. It involves exposing subjects to a visual stimulus, rapidly followed by a pattern mask, and determining whether or not this exposure has any influence on the subjects' subsequent behavior. Marcel (1983), for example, conducted a series of experiments in which subjects were subliminally exposed to a written word, and then asked to decide which of two ensuing words was either semantically or graphically similar to the initial stimulus. Marcel determined the supraliminal threshold, for each subject, by gradually reducing the onset asynchrony between stimulus and pattern mask until there was some difficulty in deciding whether or not a word had appeared. When the onset asynchrony falls below this threshold, the initial stimulus is regarded as subliminal. He found that his subjects were able to perform above chance in these forced choice judgements for stimuli between 5 and 10 msec below the supraliminal threshold. Subjects afterwards reported that they sometimes "felt silly" making a judgement about a stimulus they hadn't seen, but had simply chosen the response (in the forced choice situation) that "felt right".

Marcel takes these results to be highly significant and argues that they "cast doubt on the paradigm assumption that representations yielded by perceptual analysis are identical to and directly reflected by phenomenal percepts" (1983, p.197). Indeed, there is *prima facie* evidence here for dissociation: when a visual stimulus affects similarity judgements it is natural to assume that explicit representations have been generated by the visual system (especially when it comes to explaining successful graphical comparisons), and Marcel's results seem to indicate that this can happen without any conscious apprehension of the stimulus event. However, as usual, there are reasons to be cautious about how we interpret these results.

Holender, for example, claims that in the majority of visual masking studies an alternative interpretation of the priming effects is available, namely, that "the visibility of the primes has been much better in the priming trials than indicated by the threshold trials of these experiments" (1986, p.22). This is supported by the work of Purcell, Stewart and Stanovich (1983) who demonstrated, with respect to priming by picture, that "subjects, because of their higher level of light adaptation in the priming than in the threshold trials, were able to consciously identify the prime more often in the former than in the latter case" (Holender 1986, p.22). Holender also suggests that threshold determination may not have been adequate in a number of studies, because "when more reliable methods of threshold determination are used, semantic judgments were no better than presence-absence judgments (Nolan & Caramazza 1982)" (p.22). This issue is central to the interpretation of visual masking studies, given the statistical nature of the evidence. Indeed, Dulany has argued that "on signal detection theory, a below threshold value could still sometimes appear in consciousness and have its effect" (1991, p.109). We take the concern here, roughly speaking, to be this: a positive result in a visual masking study is a priming effect that occurs when stimulus durations are below the supraliminal threshold; but statistically significant effects only emerge within 5-10 msec of this threshold, so it's quite possible (in this stimulus-energy domain) that fluctuations in the visual system will occasionally

generate conscious events; thus, the (small) degree of priming that occurs may well be entirely due to chance conscious events.

In sum, then, it appears that the empirical evidence for dissociation is not as strong as it is often made out to be. Many of the studies we've described are methodologically flawed, in one way or another. Attempts to replicate them under more stringent conditions have often seen the relevant effects disappear, or else prove to be the result of simple, unforseen conscious processes. As a consequence it is not unreasonable to reserve judgment concerning the dissociability of explicit mental representation and phenomenal experience. This is good news for those who are attracted to vehicle theories of consciousness, since the available evidence does not appear to conclusively rule out this approach.

3 Classicism

Our next task is to determine whether either classicism or connectionism has the resources to support a vehicle theory of phenomenal consciousness. In this section we consider the various ways in which information can be represented in the brain, according to classicism, and then demonstrate why the classical approach to mental representation inevitably leads to process theories of consciousness.

3.1 Classical Styles of Mental Representation

The classical computational theory of mind holds that human cognitive processes are digital computational processes. What this doctrine actually entails about human cognition, however, is a long story, but one fortunately that is now very familiar. In a nutshell, classicism takes the generic computational theory of mind (the claim that cognitive processes are disciplined operations defined over neurally realized representational states), and adds to it a more precise account of both the representational states involved (they are complex symbol structures possessing a combinatorial syntax and semantics) and the nature of computational processes (they are syntactically-governed transformations of these symbol structures). All the rich diversity of human thought – from our most "mindless" everyday behavior of walking, sitting and opening the fridge, to our most abstract conceptual ponderings – is the result, according to the classicist, of a colossal number of syntactically-driven operations defined over complex neural symbols.[9]

Before proceeding any further, however, it is important to be clear about the entailments of this doctrine, at least as we read it. One sometimes hears it said that classicism really only amounts to the claim that human cognitive processes are *digitally simulable*: that an appropriate formalism could, in principle, reproduce the input/output profiles of our cognitive capacities. But this is a relatively weak claim; indeed, given the now standard interpretation of (what has come to be known as) the Church-Turing thesis – viz., that an appropriately constructed digital computer can, in principle at least, perform *any* well defined computational function (given enough time) – the view that human cognitive capacities can be *simulated* by digital computational processes represents nothing more than a commitment to the generic computational theory of mind.[10] Consequently, it is only under a stronger interpretation – in particular, only when it is understood as the doctrine that our cognitive processes *are* digital computational processes, and hence *are* symbol manipulations – that classicism becomes an interesting empirical thesis. What classicism requires, under this stronger interpretation, is not

---

[9] The more prominent contemporary philosophers and cognitive scientists who advocate a classical conception of cognition include Chomsky (1980), Field (1978), Fodor (1975, 1981, 1987), Harman (1973), Newell (1980), Pylyshyn (1980, 1984, 1989), and Sterelny (1990). For those readers unfamiliar with classicism, a good entry point is provided by the work of Haugeland (1981, 1985, especially Chps.2 and 3).

[10] There are, of course, substantive issues surrounding the legitimacy of this particular interpretation of Church's and Turing's original theses, but we won't buy into these here (although see Cleland 1993 and Rubel 1989 for interesting discussions).

just a formalism that captures the input/output profiles of our cognitive capacities, but, as Fodor and Pylyshyn point out, a formalism whose symbol structures are isomorphic with certain physical properties of the human brain:

> The symbol structures in a Classical model are assumed to correspond to real physical structures in the brain and the *combinatorial structure* of a representation is supposed to have a counterpart in structural relations among physical properties of the brain. For example, the relation 'part of', which holds between a relatively simple symbol and a more complex one, is assumed to correspond to some physical relation among brain states…
>
> This bears emphasis because the Classical theory is committed not only to there being a system of physically instantiated symbols, but also to the claim that the physical properties onto which the structure of the symbols is mapped *are the very properties that cause the system to behave as it does*. In other words the physical counterparts of the symbols, and their structural properties, cause the system's behavior. (Fodor & Pylyshyn 1988, pp.13-4)

In what follows, we will adopt this strong interpretation of classicism (see also Fodor 1975; and Pylyshyn 1984, 1989). Our task is to discover what this story about human cognition implies with respect to the forms of information coding in the brain.

As we pointed out in the previous section, it is commonplace for theorists to distinguish between different ways in which a computational device can carry information. Dennett (1982) has developed a taxonomy, consisting of four distinct styles of representation, that we believe respects the implicit commitments of most theorists in this area (see also Cummins 1986; and Pylyshyn 1984). We will employ this taxonomy as a useful framework within which to couch discussion of classical representation, and the prospects for a classical vehicle theory of consciousness. First, information can be represented, Dennett tells us, in an *explicit* form:

> Let us say that information is represented explicitly in a system if and only if there actually exists in the functionally relevant place in the system a physically structured object, a formula or string or tokening of some members of a system (or 'language') of elements for which there is a semantics or interpretation, and a provision (a mechanism of some sort) for reading or parsing the formula. (1982, p.216)

To take a familiar example: in a Turing machine the symbols written on the machine's tape constitute the "physically structured" vehicles of explicitly represented information. These symbols are typically subject to an interpretation (provided by the user of the machine), and can be "read" by virtue of mechanisms resident in the machine's read/write head. These symbols are thus "explicit representations", according to Dennett's taxonomy; they are physically distinct objects, each possessed of a single semantic value. In the classical context, explicit representation consists in the tokening of symbols in some neurally realized representational medium. This is a very robust form of mental representation, as each distinct item of information is encoded by a physically discrete, structurally complex object in the human brain. It is these objects upon which explicit information[11] supervenes, according to the classicist.

Dennett identifies three further styles of representation, which we'll refer to collectively as *inexplicit*. The first is *implicit* representation, defined as follows:

> [L]et us have it that for information to be represented *implicitly*, we shall mean that it is *implied* logically by something that is stored explicitly. (1982, p.216)

It is questionable, however, whether the concept of implicit representation, defined in this way, is relevant to classical cognitive science. Logical consequences don't have effects unless there are mechanisms whereby a system can *derive* (and *use*) them. And it is clear from the way Dennett

---

[11] In what follows, whenever we talk of 'explicit information' (and, shortly, of 'potentially explicit information' and 'tacit information'), this is always to be understood as a shorthand way of referring to information that is *represented* in an explicit fashion (and in a potentially explicit and tacit fashion, respectively). These more economical formulations are used purely for stylistic reasons.

defines it that implicit information can exist in the absence of such mechanisms. Another way of putting this is to say that while the information that a system implicitly represents does partly supervene on the system's physical substrate (the explicit tokens that act as premises), its supervenience base also includes principles of inference *which need not be physically instantiated*. Thus implicit representation is really just a logical notion, and not one that can earn its keep in cognitive science.

However, an implication that a system is *capable of drawing*, is a different matter. Dennett refers to information that is not currently explicit, but which a computational system is capable of rendering explicit, as *potentially explicit* (1982, pp.216-217). Representation of this form is not to be unpacked in terms of mere logical entailment, but in terms of a system's computational capacities. For example, a Turing machine is typically capable of rendering explicit a good deal of information beyond that written on its tape. Such additional information, while not yet explicit, isn't merely implicit; it is potentially explicit. And it is potentially explicit in virtue of the symbols written on the machine's tape and the mechanisms resident in its read/write head.[12]

Potentially explicit representation is crucial to classical accounts of cognition, because it is utterly implausible to suppose that everything we know is encoded explicitly. Instead, classicism is committed to the existence of highly efficient, generative systems of information storage and retrieval, whereby most of our knowledge can be readily derived, when required, from that which is encoded explicitly (i.e., from our "core" knowledge store – see, e.g., Dennett 1984; and Fodor 1987, Chp.1). In other words, on any plausible classical account of human cognition the vast majority of our knowledge must be encoded in a potentially explicit fashion. The mind has this capacity in virtue of the physical symbols currently being tokened (i.e. stored symbols and those that are part of an active process) and the processing mechanisms that enable novel symbols to be produced (data retrieval and data transformation mechanisms). Thus, in classicism, most of our knowledge is only potentially explicit. This information supervenes on those brain structures that realize the storage of symbols, and those mechanisms that allow for the retrieval, parsing and transformation of such symbols.

Dennett's taxonomy includes one further style of representation, which he calls *tacit* representation. Information is represented tacitly, for Dennett, when it is embodied in the primitive operations of a computational system (1982, p.218). He attributes this idea to Ryle:

> This is what Ryle was getting at when he claimed that explicitly proving things (on blackboards and so forth) depended on the agent's having a lot of knowhow, which could not itself be explained in terms of the explicit representation in the agent of any rules or recipes, because to be able to manipulate those rules and recipes there has to be an inner agent with the knowhow to handle those explicit items – and that would lead to an infinite regress. At the bottom, Ryle saw, there has to be a system that merely has the knowhow. If it can be said to represent its knowhow at all, it must represent it not explicitly, and not implicitly – in the sense just defined – but tacitly. The knowhow has to be built into the system in some fashion that does not require it to be represented (explicitly) in the system. (1982, p.218)

The Turing machine can again be used to illustrate the point. The causal operation of a Turing machine, remember, is entirely determined by the tokens written on the machine's tape together with the configuration of the machine's read/write head. One of the wondrous features of a Turing machine is that computational manipulation rules can be explicitly written down on the machine's tape; this of course is the basis of stored program digital computers and the possibility of a Universal Turing machine (one which can emulate the behavior of any other Turing

---

[12] Dennett tends to think of potentially explicit representation in terms of a system's processing capacity to render explicit information that is *entailed* by its explicit data. But strictly speaking, a digital system might be able to render explicit, information that is linked to currently explicit data by semantic bonds far looser than logical entailment. We count *any* information that a system has the capacity to render explicit as potentially explicit, whether or not this information is entailed by currently explicit data.

machine). But not all of a system's manipulation rules can be explicitly represented in this fashion. At the very least, there must be a set of primitive processes or operations built into the system in a non-explicit fashion, and these reside in the machine's read/write head. That is, the read/write head is so physically constructed that it behaves as if it were following a set of primitive computational instructions. Information embodied in these primitive operations is neither explicit, nor potentially explicit (since there need not be any mechanism for rendering it explicit), but tacit.

In a similar vein, tacit representation is implicated in our primitive cognitive processes, according to the classicist. These operate at the level of the symbolic atoms and are responsible for the transformations among them. No further computational story need be invoked below this level; such processes are just brute physical mechanisms. Classicists conceive them as the work of millions of years of evolution, embodying a wealth of information that has been "transferred" into the genome. They emerge in the normal course of development, and are not subject to environmental influences, except in so far as some aspects of brain maturation require the presence of environmental "triggers". So classical cognition bottoms out at symbolic atoms, implicating explicit information, and the "hardwired" primitive operations defined over them that implicate tacit information. In the classical context we can thus distinguish tacit representation from both explicit and potentially explicit styles of mental representation as follows: of the physical structures in the brain, explicit information supervenes only on tokened symbolic expressions; potentially explicit information supervenes on these structures too, but *also* on the physical mechanisms capable of rendering it explicit; in contrast to both, tacit information supervenes *only* on the brain's processing mechanisms.[13]

3.2 Classicism and Consciousness

Armed with this taxonomy of classical styles of mental representation, we can now raise the following question: Does classicism have the computational resources to support a vehicle theory of phenomenal consciousness?

Of the four styles of representation in Dennett's taxonomy, we found that only three are potentially germane to classical cognitive science, namely: explicit, potentially explicit and tacit representation (implicit representation being a merely logical notion). Consequently, a classical vehicle theory of consciousness would embrace the distinction between explicit representation (on the one hand) and potentially explicit/tacit representation (on the other), as the boundary between the conscious and the unconscious. It would hold that all phenomenal experience is the result of the tokening of symbols in the brain's representational media, and that whenever such a symbol is tokened, the content of that representation is phenomenally experienced. It would hold that whenever information is causally implicated in cognition, yet not consciously experienced, such information is encoded inexplicitly.

But, on the face of it, a classicist can't really contemplate this kind of vehicle theory of phenomenal experience. Any initial plausibility it has derives from treating the classical unconscious as a combination of both tacit and potentially explicit information, and this is misleading. Classicism can certainly allow for the *storage* of information in a potentially explicit form, but information so encoded is never *causally active*. Consider once again the operation of a Turing machine. In such a system, you'll recall, information is potentially explicit if the system has the capacity to write symbols with those contents (given the symbols currently present on its tape and the configuration of its read/write head). But while it may have this capacity, until it actually renders a piece of information explicit, this information can't influence the ongoing

---

[13] Pylyshyn's notion of the brain's "functional architecture" arguably incorporates tacit representation (1984). Both he and Fodor have been at pains to point out that classicism is *not* committed to the existence of explicit processing rules. They might *all* be hardwired into the system, forming part of its functional architecture, and it's clear that some processing rules *must* be tacit, otherwise the system couldn't operate.

behavior of the system. In fact, qua potentially explicit, such information is just as causally impotent as the logical entailments of explicit information. In order for this information to throw its weight around it must first be physically embodied as symbols written on the machine's tape. Then, and only then, when these symbols come under the gaze of the machine's read/write head, can the information they encode causally influence the computational activities of that system.

Consequently, when causal potency is at issue (rather than information coding per se) potentially explicit information drops out of the classical picture. On the classical vehicle theory under examination, this places the entire causal burden of the unconscious on the shoulders of tacit representation. Of course tacit information (unlike potentially explicit information) *is* causally potent in classical computational systems, because it's embodied in the primitive operations of such systems. Thus, an unconscious composed exclusively of tacit information would be a causally efficacious unconscious. Indeed, Pylyshyn suggests that low level vision, linguistic parsing and lexical access, for example, may be explicable merely as unconscious neural processes that "instantiate pieces of functional architecture" (1984, p.215).[14] However, despite this, it is implausible in the extreme to suppose that classicism can delegate *all* the cognitive work of the unconscious to the vehicles of tacit representation, as we'll explain.

Whenever we act in the world, whenever we perform even very simple tasks, it is evident that our actions are guided by a wealth of knowledge concerning the domain in question.[15] So in standard explanations of decision making, for example, the classicist makes constant reference to beliefs and goals that have a causal role in the decision procedure. It is also manifest that *most* of the information guiding this process is not phenomenally conscious. According to the classical vehicle theory under consideration, then, such beliefs *must be tacit*, realized as hard-wired transformations among the explicit and, by assumption, conscious states. The difficulty with this suggestion, however, is that many of the conscious steps in a decision process implicate a whole range of unconscious beliefs interacting according to unconscious rules of inference. That is, there is a complex *economy* of unconscious states that mediate the sequence of conscious episodes. While it is possible that all the *rules* of inference are tacit, the mediating train of unconscious beliefs must *interact* to produce their effects, else we don't have a causal explanation. But the only model of causal interaction available to a classicist involves explicit representations (Fodor is one classicist who has been at pains to point his out – see, e.g., his 1987, p.25). So, either the unconscious includes explicit states, or there are no plausible classical explanations of higher cognition. There seems to be no escape from this dilemma for the classicist.

There is a further difficulty for this version of classicism: it provides no account whatever of learning. While we can assume that some of our intelligent behavior comes courtesy of endogenous factors, a large part of our intelligence is a result of a long period of learning. A classicist typically holds that learning (as opposed to development or maturation) consists in the fixation of beliefs via the generation and confirmation of hypotheses. This process must be largely unconscious, since much of our learning doesn't involve conscious hypothesis testing. As above, this picture of learning requires an interacting system of unconscious representations, and, for a classicist, this means *explicit* representations. If we reject this picture, and suppose the unconscious to be entirely tacit, then there is *no* cognitive explanation of learning, in that

---

[14] Though it is worth noting that most classicists reject this picture, believing that such cognitive tasks implicate processing over intermediate explicit representations. See, e.g., Fodor 1983.

[15] This fact about ourselves has been made abundantly clear by research in the field of artificial intelligence, where practitioners have discovered to their chagrin that getting computer-driven robots to perform even very simple tasks requires not only an enormous knowledge base (the robots must know a lot about the world) but also a capacity to very rapidly access, update and process that information. This becomes particularly acute for AI when it manifests itself as the *frame problem*. See Dennett (1984) for an illuminating discussion.

learning is always and everywhere merely a process which reconfigures the brain's functional architecture. But any classicist who claims that learning is non-cognitive, is a classicist in no more than name.

The upshot of all of this is that any remotely plausible classical account of human cognition is committed to a vast amount of *unconscious symbol manipulation*. Indeed, the classical focus on the unconscious is so extreme that Fodor is willing to assert that "practically all psychologically interesting cognitive states are unconscious..." (1983, p.86). Consequently, classicists can accept that tacitly represented information has a major causal role in human cognition, and they can accept that much of our acquired knowledge of the world and its workings is stored in a potentially explicit fashion. But they cannot accept that the only explicitly represented information in the brain is that which is associated with our phenomenal experience – for every conscious state participating in a mental process classicists must posit a whole bureaucracy of unconscious intermediaries, doing all the real work behind the scenes. Thus, for the classicist, the boundary between the conscious and the unconscious cannot be marked by a distinction between explicit representation and potentially explicit/tacit representation. Whether any piece of information borne by the brain is phenomenally experienced is not a matter of whether it is encoded explicitly, but a matter of the computational processes in which it is implicated. We conclude that classicism doesn't have the computational resources required to develop a plausible vehicle theory of phenomenal consciousness. Consequently, any classicist who seeks a *computational* theory of consciousness is forced to embrace a process theory – a conclusion, we think, that formalizes what most classicists have simply taken for granted.

4 Connectionism

In this section we introduce connectionism, and show how Dennett's taxonomy of representational styles can be adapted to this alternative computational conception of cognition. This enables us to pose (and answer) a connectionist version of the question we earlier put to classicism, i.e.: Does connectionism have the computational resources to support a vehicle theory of phenomenal experience?

4.1 Connectionist Styles of Mental Representation

Whereas classicism is grounded in the computational theory underpinning the operation of conventional digital computers, connectionism relies on a neurally inspired computational framework commonly known as *parallel distributed processing* (or just PDP).[16]

A PDP network consists in a collection of processing units, each of which has a continuously variable *activation level*. These units are physically linked by connection lines, which enable the activation level of one unit to contribute to the input and subsequent activation of other units. These connection lines incorporate modifiable *connection weights*, which modulate the effect of one unit on another in either an excitatory or inhibitory fashion. Each unit sums the modulated inputs it receives, and then generates a new activation level that is some threshold function of its present activation level and that sum. A PDP network typically performs computational operations by "relaxing" into a stable pattern of activation in response to a stable array of inputs. These operations are mediated by the connection weights, which determine (together with network connectivity) the way that activation is passed from unit to unit.

The PDP computational framework does for connectionism what digital computational theory does for classicism. Human cognitive processes, according to connectionism, are the

---

[16] The *locus classicus* of PDP is the two volume set by Rumelhart, McClelland, and the PDP Research Group (Rumelhart & McClelland, 1986; McClelland & Rumelhart, 1986). Useful introductions to PDP are Rumelhart and McClelland 1986, Chps.1-3; Rumelhart 1989; and Bechtel & Abrahamsen 1991, Chps.1-4.

computational operations of a multitude of PDP networks implemented in the neural hardware in our heads. And the human mind is viewed as a coalition of interconnected, special-purpose, PDP devices whose combined activity is responsible for the rich diversity of our thought and behavior. This is the connectionist computational theory of mind.[17]

Before examining the connectionist styles of information coding, it will be necessary to clarify the entailments of this approach to cognition. There are two issues of interpretation which must be addressed, the first concerning the manner in which connectionism differs from classicism, the second concerning the relationship between PDP systems and the operation of real neural networks in the brain. We will look briefly at these in turn.

First, there has been substantial debate in recent cognitive science about the line of demarcation between connectionism and classicism. At one extreme, for example, are theorists who suggest that no such principled demarcation is possible. The main argument for this seems to be that as any PDP device can be simulated on a digital machine (in fact, the vast majority of work on PDP systems involves such simulations), it follows that connectionist models of cognition merely represent an (admittedly distinctive) subset of classical models, and hence that classicism subsumes the connectionist framework.[18] At the other extreme is a large group of theorists who insist that there is a principled distinction between these two cognitive frameworks, but nonetheless disagree with one another about its precise details.[19] We don't wish to become embroiled in this debate here. Instead, we think it suffices to point out that once one adopts the strong interpretation of classicism outlined in the previous section, the simulation argument described above loses its force: while many classicists claim that PDP represents a plausible implementation-level (i.e., non-cognitive) framework for classical models of cognition (see, e.g., Fodor & Pylyshyn 1988, pp.64-6), no classicist, as far as we know, wants to argue that the massively parallel hardware of the brain first implements a digital machine which is then employed to simulate a PDP system.[20] In what follows, therefore, we will assume that connectionism and classicism represent competing theories of human cognition.

Second, even though the PDP computational framework is clearly inspired by the neuroanatomy of the brain, there is still a substantive issue concerning the exact relationship between PDP systems and the operation of real neural networks. Connectionists are divided on this issue. On the one hand, theorists such as Rumelhart and McClelland have been explicit about the fact that PDP systems directly model certain high-level physical properties of real neural networks. Most obviously, the variable activation levels of processing units and the modifiable weights on connection lines in PDP networks directly reflect the spiking frequencies of neurons and the modulatory effects of synaptic connections, respectively (see, e.g., Rumelhart

---

[17] Some of the more prominent contemporary philosophers and cognitive scientists who advocate a connectionist conception of cognition include Clark (1989, 1993), Cussins (1990), Horgan and Tienson (1989), Rumelhart and McClelland (Rumelhart & McClelland, 1986; McClelland & Rumelhart, 1986), Smolensky (1988), and the earlier Van Gelder (1990). For useful introductions to connectionism, see Bechtel & Abrahamsen 1991; Clark 1989, Chps.5-6; Rumelhart 1989; and Tienson 1987.

[18] We say that this is the main argument for this deflationary interpretation of connectionism, but it's hard to find any explicit formulation in published work, though one certainly comes across it in email discussions of these issues.

[19] Each of the following theorists, for example, provides a somewhat different account of how this distinction ought to be characterized: Bechtel (1988a), Cussins (1990), Fodor and Pylyshyn (1988), Hatfield (1991), Horgan and Tienson (1989), O'Brien (1993), and Smolensky (1988).

[20] In this context, the fact that PDP networks can be simulated on digital equipment is not much more significant than the fact that, say, meteorological phenomena can. The only real difference is that in the former case, but not the latter, one computational device is being employed to simulate the activity of another. In both cases, though, real properties of the phenomenon being simulated are missing. These properties are very obvious in the case of the weather. In the case of the simulation of PDP systems, on the other hand, the omissions are more subtle. One such property is real time performance. Another, we shall argue, is phenomenal experience. But more on this later (see Section 5.1).

& McClelland 1986, Chp.4). Sejnowski goes even further, arguing that while PDP systems do not attempt to capture molecular and cellular detail, they are nonetheless "stripped-down versions of real neural networks similar to models in physics such as models of ferromagnetism that replace iron with a lattice of spins interacting with their nearest neighbors" (1986, p.388). Smolensky, on the other hand, argues that because we are still largely ignorant about the dynamical properties of the brain that drive cognitive operations, and because the PDP framework leaves out a number of properties of the cerebral cortex, a proper treatment of connectionism places it at a level once removed from real neural networks (1988).

Our own interpretation of the relationship between PDP systems and real neural networks locates us at the former end of this spectrum (see also Bechtel 1988b and Lloyd 1988). Like Sejnowski, we think that the PDP computational framework is best understood as an *idealized* account of real neural networks. As with any idealization in science, what goes into such an account depends on what properties of neural nets one is trying to capture. The idealization must be complex enough to do justice to these properties, and yet simple enough that these properties are sufficiently salient (see, e.g., Churchland & Sejnowski 1992, Chp.3). In this respect, the PDP framework isolates and hence enables us to focus on the *computationally* significant properties of neural nets, while ignoring their fine-grained neurochemistry. Our best neuroscience informs us that neural nets compute by generating patterns of neural activity in response to inputs, and that these patterns of activity are the result of the modulatory effects of synapses in the short-term, and modifications to these synapses over the longer term. It's precisely these structural and temporal properties that are captured by the networks of processing units and connection weights that comprise PDP systems. Of course, there are all sorts of details in the current specification of the PDP framework that are likely to prove unrealistic from the biological perspective (the back propagation learning procedure is an oft-cited example – see, e.g., the discussion in Churchland & Sejnowski 1992, Chp.3). But this does not impugn the integrity of the framework as a whole. What's more, it is entirely open to connectionists to incorporate more complex dynamical features of neural nets, if these are subsequently demonstrated to be crucial to the computational operation of the brain.

One final point is in order, in this context. It is crucial to distinguish between the PDP computational framework itself (as generically described in the preceding paragraphs), and the "toy" PDP models of (fragments of) human cognitive capacities that one can find in the literature (Sejnowski and Rosenberg's NETtalk (1987), which learns to transform graphemic input into phonemic output, is a much discussed example). In interpreting the former as an idealized account of the operation of real neural networks, we don't mean to suggest that the latter models are in any way biologically realistic. These toy models are interesting and important because they demonstrate that even very simple networks of processing units (simple, at least, when compared with the complexity and size of real neural networks) can realize some powerful information processing capacities. But it would clearly be implausible to suppose that such models describe the manner in which these cognitive capacities are actually realized in human brains. What is not so implausible is that these models capture, albeit in a rudimentary way, the *style* of computation that is employed by the brain's own neural networks.

With these issues of interpretation behind us, it is now time to consider what the connectionist conception of human cognition entails about the way information is encoded in the brain. While it was formulated in the context of digital computational theory, Dennett's taxonomy is also applicable to the PDP framework (and hence to connectionism), because there are connectionist analogues of explicit, potentially explicit, and tacit styles of representation, as we shall now demonstrate.

The representational capacities of PDP systems rely on the plasticity of the connection weights between the constituent processing units.[21] By altering these connection weights, one alters the activation patterns the network produces in response to its inputs As a consequence, an individual network can be taught to generate a range of stable target patterns in response to a range of inputs. These stable patterns of activation are semantically evaluable, and hence constitute a transient form of information coding, which we will refer to as *activation pattern representation*.

In terms of the various styles of representation that Dennett describes, it is reasonable to regard the information encoded in stable activation patterns across PDP networks as *explicitly* represented. For these patterns are physically discrete, structurally complex objects, which, like the symbols in conventional computers, each possess a single semantic value – no activation pattern ever represents more than one distinct content. These stable patterns are embedded in a system with the capacity to process them in structure sensitive ways. An activation pattern is "read" in virtue of having effects elsewhere in the system. That is why stability is such a crucial feature of activation pattern representations. Being stable enables an activation pattern to contribute to the clamping of inputs to other networks, thus generating further regions of stability (and ultimately contributing to coherent schemes of action). Moreover, the quality of this effect is *structure sensitive* (ceteris paribus), that is, it is dependent on the precise profile of the source activation pattern. While the semantics of a PDP network is not language-like, it typically involves some kind of systematic mapping between locations in activation space and the object domain.[22]

While activation patterns are a transient feature of PDP systems, a "trained" network has the capacity to generate a whole range of activation patterns, in response to cueing inputs. So a network, in virtue of its connection weights and pattern of connectivity, can be said to *store* appropriate responses to input. This form of information coding, which is sometimes referred to as *connection weight representation,* constitutes long-term memory in PDP systems. This long-term storage of information is *superpositional* in nature, since *each* connection weight contributes to the storage of *every* stable activation pattern (every explicit representation) that the network is capable of generating. Consequently, the information that is stored in a PDP network is not encoded in a physically discrete manner. The one appropriately configured network encodes a *set* of contents corresponding to the range of explicit tokens it is disposed to generate. For all these reasons, a PDP network is best understood as storing information in a *potentially explicit* fashion. This information consists of all the data that the network has the capacity to render explicit, given appropriate cueing inputs.

Finally, what of *tacit* representation? You'll recall that in the conventional context tacit information inheres in those primitive computational operations (defined over symbolic atoms) that are hardwired into a digital computer. In the PDP framework the analogous operations depend on the individual connection weights and units, and consist in such processes as the modulation and summation of input signals and the production of new levels of activation. These operations are responsible for the generation of explicit information (stable patterns of activation) within PDP networks. It is natural to regard them as embodying tacit information, since they completely determine the system's response to input.

---

[21] For good general introductions to the representational properties of PDP systems, see Bechtel & Abrahamsen 1991, Chp.2; Churchland 1995; Churchland & Sejnowski 1992, Chp.4; Rumelhart & McClelland 1986, Chps.1-3; and Rumelhart 1989. More fine-grained discussions of the same can be found in Clark 1993; and Ramsey, Stich & Rumelhart 1991, Part II.

[22] Here we are relying on what has become the standard way of distinguishing between the explicit representations of classicism and connectionism, whereby the former, but not the latter, are understood as possessing a (concatenative) combinatorial syntax and semantics. The precise nature of the internal structure of connectionist representations, however, is a matter of some debate; see, e.g., Fodor & Pylyshyn 1988; Smolensky 1987; and Van Gelder 1990.

4.2 Connectionism and Consciousness

With these PDP styles of representation before us, let's now address the key question: Does connectionism have the computational resources to support a vehicle theory of consciousness? Just as was the case with the classical version, such a connectionist vehicle theory would embrace the distinction between explicit representation and potentially explicit/tacit representation, as the boundary between the conscious and the unconscious. It would hold that each element of phenomenal experience corresponds with the generation of an activation pattern representation somewhere in the brain, and conversely, that whenever such a stable pattern of activation is generated, the content of that representation is phenomenally experienced. Consequently, this connectionist vehicle theory would hold that whenever unconscious information is causally implicated in cognition, such information is not encoded in the form of activation pattern representations, but merely inexplicitly, in the form of potentially explicit/tacit representations.

But is this suggestion any more plausible in its connectionist incarnation, than in the classical context? We think it is. In the next section we'll develop this suggestion in some detail. For now, we merely wish to indicate which features of PDP-style computation make this connectionist vehicle theory of consciousness worth considering, even though its classical counterpart is not even remotely plausible.

While we were able to apply Dennett's taxonomy to both classicism and connectionism, there is nonetheless an important representational asymmetry between these two competing theories of cognition. Whereas potentially explicit information is causally impotent in the classical framework (it must be rendered explicit before it can have any effects), the same is not true of connectionism. This makes all the difference. In particular, whereas classicism, using only its inexplicit representational resources, is unable to meet all the causal demands on the unconscious (and is thus committed to a good deal of unconscious symbol manipulation), connectionism holds out the possibility that it can (thus leaving stable activation patterns free to line up with the contents of consciousness).

Potentially explicit information is encoded in a PDP network in virtue of its relatively long-term capacity to generate a range of explicit representations (stable activation patterns) in response to cueing inputs. This capacity is determined by its configuration of connection weights and pattern of connectivity. However, we saw earlier that a network's connection weights and connectivity structure is also responsible for the manner in which it responds to input (by relaxing into a stable pattern of activation), and hence the manner in which it processes information. This means that the causal substrate driving the computational operations of a PDP network is *identical* to the supervenience base of the network's potentially explicit information. So there is a strong sense in which it is the potentially explicit information encoded in a network (i.e., the network's "memory") that actually governs its computational operations.

If potentially explicit information governs the computational operations of a PDP network, what becomes of the distinction between potentially explicit and tacit representation? For all practical purposes the distinction lapses, since, in PDP systems, potentially explicit and tacitly represented information *have the same supervenience base.* (This is another way of expressing the oft-cited claim that connectionism dispenses with the classical code/process distinction – see, e.g., Clark 1993). As a consequence, tacitly represented information, understood as the information embodied in the primitive computational operations of the system, *is identical to* potentially explicit information, understood as the information that the system has the capacity to render explicit.

This fact about PDP systems has major consequences for the manner in which connectionists conceptualize cognitive processes. Crucially, information that is merely potentially explicit in PDP networks need not be rendered explicit in order to be causally efficacious. There is a real sense in which *all* the information that is encoded in a network in a

potentially explicit fashion is causally active *whenever* that network responds to an input. What is more, learning, on the connectionist story, involves the progressive modification of a network's connection weights and pattern of connectivity, in order to encode *further* potentially explicit information. Learning, in other words, is a process which actually reconfigures the potentially-explicit/tacit representational base, and hence adjusts the primitive computational operations of the system. In Pylyshyn's (1984) terms, one might say that learning is achieved in connectionism by modifying a system's functional architecture.

The bottom line in all of this is that the inexplicit representational resources of connectionist models of cognition are vast, at least in comparison with their classical counterparts. In particular, the encoding and, more importantly, the *processing* of acquired information, are the preserve of causal mechanisms that don't implicate explicit information (at least, not until the processing cycle is complete and stable activation is achieved). Consequently, most of the computational work that a classicist must assign to unconscious symbol manipulations, can in connectionism be credited to operations implicating inexplicit representation. Explicit representations, on this alternative conception, are the *products* of unconscious processes, and thus a connectionist can feel encouraged in the possibility of aligning phenomenal experience with these representational vehicles.

Connectionism, while remaining a *computational* conception of cognition, paints a cognitive landscape quite distinct from its classical counterpart. In summary the connectionist story goes something like this. Conscious experiences are stable states in a sea of unconscious causal activity. The latter takes the form of *intra*-network "relaxation" processes, that result in stable patterns of activation, and which are determined by the superpositionally encoded information stored therein. Unconscious processes thus *generate* activation pattern representations, which the connectionist is free to identify with individual phenomenal experiences, since none is required to account for the unconscious activity itself. The unconscious *process*, entirely mediated by superpositionally encoded data, generates a conscious *product*, in the form of stable patterns of activation in neurally realized PDP networks.

So connectionism does appear to have the right computational profile to hazard a vehicle theory of consciousness. Since such theories are all but absent from contemporary cognitive science, we feel it is worth exploring this much neglected region of the theoretical landscape. In the next section we do just this by providing a sketch of a connectionist theory that identifies phenomenal experience with the brain's generation of explicit representations. We believe that once this account is laid bare, and some initially counter-intuitive features defended, it appears as a robust, insightful and defensible alternative to the plethora of process theories in the literature.

5 A Connectionist Vehicle Theory of Phenomenal Experience

A vehicle theory of consciousness holds that phenomenal experience is to be explained, not in terms of what explicit mental representations *do*, but in terms of what they *are*. Connectionism, we have argued, has the representational resources to venture such a theory of phenomenal consciousness. Given the power of the connectionist styles of inexplicit representation to account for unconscious thought processes and learning, it is possible to align phenomenal experience with *explicit* information coding in the brain. And that, baldly stated, is the connectionist vehicle theory of consciousness we wish to defend: phenomenal experience is identical to the brain's explicit representation of information, in the form of stable patterns of activation in neurally realized PDP networks. This amounts to a simple, yet bold empirical hypothesis, with testable consequences. In this section we develop this hypothesis in some detail by considering it both at the level of individual neural networks (the *intra*-network level) and at the higher level of the brain's global architecture (the *inter*-network level). We then finish with some very brief remarks

about how this conjecture contributes a solution to the so-called "hard" problem of phenomenal consciousness (Nagel 1974; Chalmers 1995, 1996).

5.1 The Intra-Network Level

The connectionist account of consciousness we have proposed is not completely novel. Theorists involved in laying the foundations of the connectionist approach to cognition recognized a potential role for stable patterns of activation in an account of phenomenal experience. In the very volumes in which connectionism receives its first comprehensive statement (Rumelhart & McClelland 1986; McClelland & Rumelhart 1986), for example, we find the suggestion that:

> …the contents of consciousness are dominated by the relatively stable states of the [cognitive] system. Thus, since consciousness is on the time scale of sequences of stable states, consciousness consists of a sequence of interpretations – each represented by a stable state of the system. (Rumelhart, Smolensky, McClelland & Hinton 1986, p.39)

And in another seminal piece, Smolensky makes a similar suggestion:

> The contents of consciousness reflect only the large-scale structure of activity patterns: subpatterns of activity that are extended over spatially large regions of the network and that are stable for relatively long periods of time. (1988, p.13)

It is worth pointing out, however, that neither Rumelhart, Smolensky, McClelland and Hinton, nor Smolensky, take the presence of a stable pattern of activation to be both necessary and sufficient for consciousness. Rumelhart et al. don't appear to regard stability as *necessary* for consciousness, for they suppose "that there is a relatively large subset of total units in the system whose states of activity determine the contents of consciousness", and that "the time average of the activities of these units over time periods on the order of a few hundred milliseconds correspond to the contents of consciousness" (1986, p.39). But this implies that "on occasions in which the relaxation process is especially slow, consciousness will be the time average over a dynamically changing set of patterns" (1986, p.39). In other words, stability is not *necessary* for conscious experience, since even a network that has not yet stabilized will, on this account, give rise to some form of consciousness. Smolensky, on the other hand, doesn't regard stable activation to be *sufficient* for consciousness, and says as much (1988, p.13). Consequently, it is not clear that either of these early statements actually seeks to *identify* consciousness with stable activation patterns in neurally realized PDP networks, as we are doing.

More recently, Mangan (1993a, 1996) has argued for what we are calling a vehicle theory of phenomenal experience; consciousness, he tells us, is a species of "information-bearing medium", such that the transduction of information into this special medium results in it being phenomenally experienced (see also Cam 1984; Dulany 1996). What is more, Mangan regards connectionism as a useful source of hypotheses about the nature of this medium. In particular, he suggests that the kind of approach to consciousness developed by Rumelhart, Smolensky, McClelland and Hinton can be used to accommodate vague, fleeting and peripheral forms of experience (what, following William James (1890), he calls the "fringe" of consciousness) within a computational framework (see Mangan 1993b). But like Rumelhart et al., Mangan seems to accept the possibility that states of consciousness could be associated with networks that have not fully stabilized – i.e., with *stabilizing* networks – rather than restricting them to stable patterns of activation across such networks.

Finally, Lloyd (1991, 1995 and 1996) comes closest to advancing the kind of connectionist vehicle theory of consciousness that we advocate. Recognizing the need for a principled distinction between conscious and unconscious cognition he makes the following proposal:

> Vectors of activation…are identical to conscious states of mind. The cognitive unconscious, accordingly…[consists] of the rich array of dispositional capacities latent in the weights or connection strengths of the network. (1995, p.165)

Lloyd provides a detailed analysis of phenomenal experience, developing the distinctions between sensory and non-sensory, primary and reflective forms of consciousness. He goes on to show how, on the basis of the identity claim above, these various distinctions can be cached out in connectionist terms (1995, 1996). But, again, Lloyd appears to focus his efforts on activation patterns in general, rather than stable patterns of activity, and so his account in this respect is still at some variance with ours.[23]

Why then have we made *stability* such a central feature of our connectionist account? The answer is quite straightforward: only stable patterns of activation are capable of encoding information in an explicit fashion in PDP systems, and hence only these constitute the vehicles of explicit representation in this framework. Prior to stabilization, the activation levels of the constituent processing units of a PDP network are rapidly changing. At this point in the processing cycle, therefore, while there certainly is plenty of activity across the network, there is no determinate *pattern* of activation, and hence no single, physically structured object that can receive a fixed interpretation. A connectionist vehicle theory of consciousness is thus committed to identifying phenomenal experience with *stable* patterns of activation across the brain's neural networks. On this story, a conscious experience occurs whenever the activity across a neural network is such that its constituent neurons are firing simultaneously at a *constant* rate. The physical state realized by this network activity, the complex physical object constituted by the stable pattern of spiking frequencies, *is* the phenomenal experience.

There are a couple of points that are worth making in passing here. The first is that the existence of stable patterns of activation at the level of neural networks is quite consistent with the *seamless* nature of our ongoing phenomenal experience. This is because such stabilizations can occur very rapidly; given their chemical dynamics, it's possible for real neural networks to generate many stable states per second (Churchland & Sejnowski 1992, Chp.2). Consequently, what at the level of an individual neural network is a rapid *sequence* of stable patterns, may at the level of consciousness be a *continuous* phenomenal stream.

The second is that, considered as a complex physical object, the stable activation pattern is absent in digital simulations of PDP systems. In such simulations, the activation values that compose a network's activation pattern are typically recorded in a complex array, each of whose elements is subject to updating according to the algorithms that model the network's activity. But this data structure is not equivalent to a pattern of activation across a real (non-simulated) PDP network. The latter is an object constructed from physically connected elements (such as neurons), each of which realizes a continuously variable physical property (such as a spiking frequency) of a certain magnitude. The former, by contrast, is a *symbolic representation* of such an object, in that it consists of a set of discrete symbol structures that "describes" in a numerical form the individual activation levels of a network's constituent processing units. An activation pattern across a real network thus has a range of complex structural properties (and consequent causal powers) that are not reproduced by the data structures employed in simulations. This fact is most vividly demonstrated by the temporal asymmetries that exist between real PDP networks and their digital simulations: the simulations are notoriously slow at processing information, when compared to their real counterparts, in spite of the incredible computational speed of the digital machines on which they are run. The bottom line here is that a simulated stable pattern of activity is no more a stable activation pattern than a simulated hurricane is a hurricane. Consequently, because stable patterns of activity are absent in digital simulations of PDP systems, so are phenomenal experiences, on our account.

---

[23] Lloyd recently appears to have retreated somewhat from his bold initial position. It is possible, he tells us, "to identify conscious states of mind with the hidden layer exclusively…" (1996, p.74). This move relegates activation patterns over the input layer to the status of "an underlying condition for sensory consciousness" (p.74), thus limiting his identity hypothesis to a particular subclass of the activation patterns present in neurally realized PDP networks.

There are further reasons to focus on *stable* patterns of activity, when thinking about phenomenal consciousness, rather than network activity more generally. One of these is that neurons in the brain, when not subject to inputs, fire spontaneously at random rates (Churchland & Sejnowski 1992, p.53). Consequently, there is "activity" across the neural networks of the brain, even in dreamless sleep. But, clearly, this activity doesn't produce any conscious awareness. Why not? On the connectionist vehicle theory we are proposing the answer is simple: while there is neural activity, no stable patterns of activation are generated. Of course, the neural networks of *dreaming* subjects are not active in a merely random fashion, but equivalently such subjects are not phenomenally unconscious. On our account, dreams, just like normal waking experiences, are composed of stable patterns of activity across these networks.

Another reason for focusing on stable activation patterns is one we mentioned in the previous section when introducing this style of representation. We noted there that only stable patterns of activation can facilitate meaningful communication *between* PDP networks, and hence contribute to coherent schemes of action. In PDP systems such effects are mediated by the flow of activation along connection lines, and its subsequent integration by networks downstream. No network can complete its processing (and thereby generate explicit information) unless its input is sufficiently stable. But stable input is the result of stable output. Thus, one network can contribute to the generation of explicit information in another only if itself in the grip of an explicit token. The message is: stability begets stability.

It is important to be aware, however, that in emphasizing the information processing relations enjoyed by these explicit representational states, we are not claiming that these vehicles must have such effects in order for their content to be phenomenally experienced. This, of course, would amount to a process theory of consciousness. On the vehicle theory we have been developing, phenomenal experience is an intrinsic, physical, *intra-network* property of the brain's neural networks. On this account, therefore, *inter-network information processing relations depend on phenomenal experience, not the reverse*.[24] Moreover, the presence of phenomenal experience is necessary, but not sufficient, for such inter-network communications. Explicit tokenings are not guaranteed to have information bearing effects between networks, because such effects are also contingent on the pattern of connectivity and the degree of modularity that exists in the system (not to mention the possibility of pathological failures of access). Thus while phenomenal consciousness facilitates such information processing relations, it can exist in their absence.

We have started to talk about the important role stable patterns of activation play in inter-network information processing. This is a much neglected region in connectionist theorising, most of which tends to focus on intra-network activity[25]. In the next subsection we want to partially redress this deficiency by considering the picture of consciousness that is painted by our connectionist vehicle theory at this more global level of description.

5.2 The Inter-Network Level

Theorists sometimes construe connectionism as the claim that the mind is a single, extremely complex network, and consequently find it tempting to attribute network-level properties to the mind as a whole. But this is surely a mistake. Many lines of evidence suggest that there's a significant degree of modularity in brain architecture. Connectionism is constrained by this

---

[24] This intimate relationship between inter-network information processing relations and phenomenal experience, partially explains the popularity of process theories which hold that those mental contents are conscious whose explicit vehicles have *rich* and *widespread* informational effects in a subject's cognitive economy (e.g., Baars 1988; and Dennett 1991). Since such information processing relations are always associated with phenomenology, it is tempting to suppose that it is rich and widespread informational effects that *constitute* consciousness. But, assuming our account, this is to put the cart before the horse: there is no path leading from information bearing effects to consciousness; consciousness precedes, and is responsible for, such effects.

[25] Important exceptions here are Clark & Karmiloff-Smith 1993; and Clark & Thornton forthcoming.

evidence, and so treats the mind as a large collection of interconnected, specialized PDP networks, each with its own connectivity structure and potential patterns of activity. This implies that from moment to moment, as the brain simultaneously processes parallel streams of input, and ongoing streams of internal activity, a large number of stable patterns of activation are generated across hundreds (perhaps even thousands) of neural networks. In other words, according to connectionism, from moment to moment the brain simultaneously realizes a large number of explicit representations.

This feature of connectionism has important implications for the theory of consciousness we are proposing. According to that theory, each explicit representation – each stable activation pattern – is identical to a phenomenal experience. In particular, each explicit representation is identical to an experience in which the information content encoded by that explicit vehicle is "manifested" or "displayed" – that is, the "what-it-is-likeness" of each phenomenal experience is constituted by the information content that each explicit representation encodes. But since connectionism holds that there are many such representations being tokened at each instant, the connectionist vehicle theory of consciousness implies that instantaneous phenomenal experience is in fact a very complex aggregate state composed of a large number of distinct phenomenal elements. Moreover, since the neural vehicles of explicit representation are thought to be very numerous, the connectionist vehicle theory of consciousness also implies that the neurological basis of consciousness is manifold, i.e., that there are a multitude of consciousness-making mechanisms in the brain.

While there are those who might be prepared to reject one or other of these implications, we suggest that they are quite consistent with the existing evidence, both phenomenological and neurological. Consider first the evidence of experience. Even the most casual inspection of your moment by moment phenomenal experience reveals it to be a very complex affair. Right now, as you concentrate on understanding the type-written sentences before you, your (global) phenomenal experience is simultaneously multi-modal and multi-channelled: visual experiences (the shape and color of the words on the page), language-understanding experiences (what the words and sentences mean), auditory experiences (noises drifting into the room in which you sit), tactile experiences (the chair pressing against your body), proprioceptive experiences (the position of your limbs) and so forth, *together* comprise your instantaneous *phenomenal field*. And when, for example, you *visually experience* these words, the other aspects of your phenomenal field don't momentarily disappear: you don't stop feeling where your limbs are; you don't stop having auditory experiences; you don't stop feeling the chair pressing against your lower body. In other words, instantaneous consciousness is a *polymodal composite* – a sum of concurrent, but distinct phenomenologies.[26]

To reiterate: instantaneous consciousness is not restricted to a single modality at a time. It is a complex amalgam of many contents, which, for the most part, are so constant that it's easy to take them for granted. We know of the persistence of visual experience, for instance, because we are all familiar with the decrement in phenomenology that accompanies closing our eyes. But people must often suffer severe neurological damage before they can even acknowledge the

---

[26] Some will object to these claims on the grounds that consciousness is co-extensive with attention, and attention is clearly restricted to a single focal object at a time. However, this strikes us as a mistaken view of the relationship between consciousness and attention. Attention serves to heighten some aspects of experience over others; it moves like a searchlight through the phenomenal field, but it doesn't define that field – there is plenty of phenomenology that falls outside its beam.

This still leaves us in need of some account of attention. A proponent of the connectionist vehicle theory of consciousness might attempt to explain attention in terms of mechanisms that subject information already extracted from the world, and hence already displayed in the phenomenal field, to more intense processing. Such additional processing would require the engagement of extra neural networks, which in generating further stable patterns of activation would produce an enhanced or augmented phenomenal experience of the aspect of the world in question. Jackendoff develops a similar – though not specifically connectionist – account of attention (see his 1987, pp.280-3).

existence of other persisting aspects of this field. For example, Sacks describes the tragic case of a woman who, due to acute polyneuritis of the spinal and cranial nerves throughout the neuraxis, suddenly loses her capacity to have proprioceptive experiences: "Something awful's happened," she tells Sacks, "I can't feel my body. I feel weird – disembodied" (1985, p.44). This woman has *none* of the usual (proprioceptive) feedback from her body. Without it she recognizes (perhaps for the first time) what she had, but has now lost: the feeling of embodiment. Most of us don't realize that we *don't* feel disembodied, but she is in the horrible position of having this realization forced upon her. The experience of embodiment is a constant feature of our phenomenal field.

Having said this, it is important to recognize that the various modes of experience are relatively independent of one another. Total deficits in sight and audition are quite common, and can be brought on suddenly by localized damage which leaves the other modalities more or less intact. They are like so many strands in a woven cloth – each strand adds to the cloth, but, since they run side by side, the loss of any one strand doesn't deform or diminish the others, it merely reduces the total area of fabric.

This independence among the parts of experience is even evident, to some extent, *within*

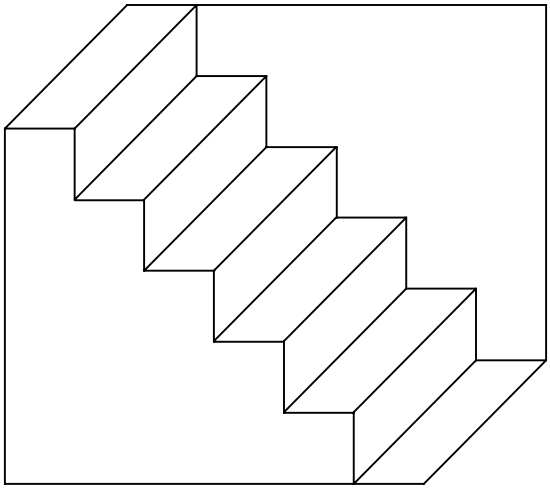

Figure 1. Inverting stairs ambiguous figure.

modalities. Consider the familiar "inverting stairs" ambiguous figure (Figure 1). It can be seen as a flight of stairs in normal orientation, with rear wall uppermost; as an inverted flight of stairs, with front wall uppermost; or even as a flat line drawing, with no perspective. And whichever of these interpretations one adopts, the details of line and space remain the same. That is, our experience here incorporates not only lines, and regions, but also some abstract phenomenology (in this case, *a sense of perspective*), phenomenology which is subject to a degree of voluntary control. Or consider the "vase/faces" ambiguous figure (Figure 2). Whether one interprets it as a vase (dark figure, light background), or as a pair of faces (light figure, dark background), there is no change in the experience of tone and line itself. Again there is some primary visual experience (i.e., the experience of lines, boundaries, light and dark regions), to which an additional variable element of abstract phenomenology is added (in this case, *object recognition*). What is striking in both these cases is the looseness of fit between the more abstract and the more concrete parts of experience.

The real force of this phenomenological evidence only fully emerges when it is conjoined with the available neuroscientific evidence. We know, on the basis of deficit studies, that the information processing that supports conscious experience is realized in structures distributed right across the brain. And the distributed nature of this information processing is both an intra-

modal and an inter-modal affair. Consider, again, our visual experience. Recent work in the neurosciences has shown that visual processing is highly modularized; the visual cortex appears to contain separate subsystems for the processing of information about color, shape, depth and even motion. When any one of these subsystems is damaged, the particular element of visual experience it supports drops out, more or less independently of the others. Take motion perception for example. Zeki relates the case of a woman who, due to a vascular disorder in the brain which resulted in a lesion to a part of the cortex outside the primary visual area, lost the ability to detect motion visually. This was so severe that,

> She had difficulty, for example, in pouring tea or coffee into a cup because the fluid appeared to be frozen, like a glacier. In addition, she could not stop pouring at the right time since she was unable to perceive the movement in the cup (or a pot) when the fluid rose. The patient also complained of difficulties in following a dialogue because she could not see the movement of…the mouth of the speaker. (Quoted in Zeki 1993, p.82)

Zeki notes that this was not a total defect in the appreciation of movement "because the perception of movement elicited by auditory or tactile stimulation was unaffected" (p.82). Moreover, her perception of other visual attributes appeared to be normal. Similarly striking case studies are available in relation to the loss of color sensations (see, for example, Sacks 1995, pp.1-38).

Deficit studies like these contain two messages. First, they confirm the picture of consciousness as an aggregate of relatively independent parts, because they demonstrate total

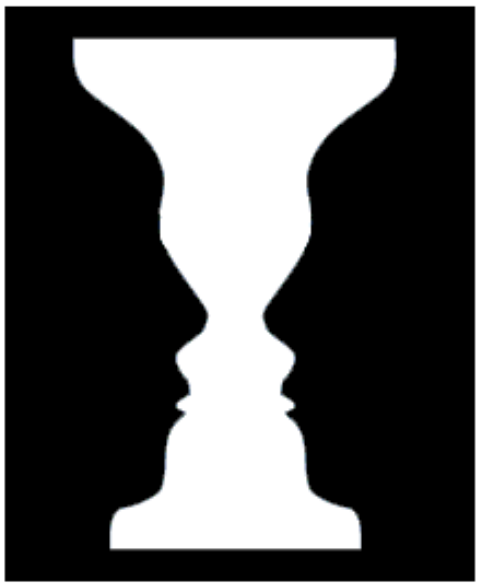

Figure 2. Vase/Faces ambiguous figure.

experiences in which one or other of the usual phenomenal elements has been subtracted. Second, they suggest a very natural way of interpreting the patently distributed nature of brain-based information processing: as evidence for the multiplicity of consciousness-making mechanisms in the brain. For it is not just cognitive capacities that are effaced as a result of cortical lesions – there are corresponding deficits and dissociations in experience. Given that

such deficits are so tightly correlated with damage to particular regions of the brain, the most parsimonious story to be told is that consciousness is generated locally at these very sites.[27]

But if our instantaneous phenomenal field is a complex amalgam of distinct and separable phenomenal elements, and if the most reasonable construal of the available neurological evidence is that there is a multiplicity of consciousness-making mechanisms distributed across the brain, then the connectionist vehicle theory of consciousness is just the sort of account we need. On this account, phenomenal experience has the complex synchronic structure it does precisely because it consists of a multitude of physically distinct explicit representations generated across the brain from moment to moment. And on this account, phenomenal experience exhibits patterns of breakdown consistent with a high degree of neural distribution because the very mechanisms that fix explicit contents in the brain are those that generate consciousness.

In addition to its capacity to account for the composite nature of phenomenal experience, the connectionist vehicle theory we advocate offers an approach to another important feature of consciousness, namely, the varying degrees of abstractness displayed by its elements. This territory is sometimes negotiated with the distinction between sensory and non-sensory kinds of experience (Lloyd 1996). According to Lloyd, sensory experiences, unlike non-sensory experiences, are modality specific; basic (meaning that they are not constituted by or dependent on other elements or experience); relatively few in number; and compulsory (pp.65-7). Of course, what is being marked here are the ends of a continuum. There are many subtle gradations along the dimensions Lloyd proposes, leading from very basic, modality-dependent elements of experience, to phenomenal elements that are more or less independent of a particular modality, but are decidedly non-basic. Before explaining how this continuum emerges quite naturally from the connectionist vehicle theory of consciousness, it will therefore be useful to take a further brief survey of phenomenal experience.

We introduced the idea that consciousness incorporates elements of varying degrees of abstractness in relation to the figures described above. The phenomenology of each of these figures incorporates, in addition to the more concrete experience of line and tone, a perspectival or figurative element (a gestalt) which is demonstrably distinct from its concrete ground (see above). Even the experience of depth in binocular vision is to some extent more abstract than other elements of the visual field. It can be removed simply by shutting one eye. Most scenes then loses something – a quality of extension let's say – which returns immediately upon opening the closed eye (try this with a set of exposed beams in a ceiling, or a row of books along a bookshelf). The point here is that depth perception is something added to basic visual experience – one can have rich and informative visual experience without it – yet it is a genuine part of the phenomenology when it is present (there is "something it is like" to perceive depth).

A further example of this kind concerns the recognition of faces. Humans are supremely good both at remembering faces, and at noticing a familiar visage in a crowd of passing strangers. This capacity is something above and beyond the mere ability to perceive faces (a stranger's face is no less a face for its lack of familiarity) and has its own accompanying phenomenology – there is "something it is like" to recognize a familiar face. Note that this case is slightly different from the gestalt experiences described above, because we are here describing

---

[27] There are echoes here of Dennett's multiple drafts theory of consciousness (1991, 1993). Dennett, like us, resists the idea that there is a single stream of consciousness, claiming that there are instead "multiple channels in which specialist circuits try, in parallel pandemoniums, to do their various things, creating Multiple Drafts as they go" (1991, pp.253-254). He further rejects what he calls the "Cartesian theatre" model of consciousness; the idea that there is a single structure or system in the brain where the contents of consciousness all come together for the delectation of the mind's eye. Consciousness, instead, is the result of processes (Dennett calls them "microtakings") distributed right across the brain. For more neuro-psychological evidence pointing to the distributed neural basis of consciousness, see, e.g., the papers in Milner & Rugg 1992.

an element of experience further to the mere perception of a face as an organized whole. A familiar face is not only perceived as a face, but as a face with a familiar "feel". This "feeling of familiarity" (see Mangan 1993b for further discussion) is superordinate to facial perception *simpliciter*.[28] It is also to be distinguished from the capacity to associate a name with a face. For those who have difficulty recalling names, the feeling of familiarity on meeting a casual acquaintance often arises (with great embarrassment) well before that person's name returns.

A particularly important kind of abstract experience arises, among other places, in the context of speech perception. The sounds we use to communicate appear to be subject to a whole series of processing stages prior to the emergence of their meanings. The sonic stream must be segmented into phonemes, then morphemes (the smallest units of meaning), words, phrases and sentences. These various processes generate phenomenal elements of varying degrees of abstractness, from basic sound elements, through word and phrase gestalts, and culminating in what Strawson (1994, pp.5-13) calls "understanding experience". On the latter, consider the difference between Jacques (a monoglot Frenchman) and Jack (a monoglot Englishman) as they listen to the news in French. (This example comes from Strawson 1994, pp.5-6.) While there is a sense in which Jacques and Jack have the same aural experience, their experiences are utterly different in another respect. Jacques *understands* what he hears, while Jack does not. This difference is not just a difference in Jacques' capacity to respond to what he hears, it is a difference *within* phenomenal experience. Jacques consciously experiences something that Jack does not. Understanding-experience is that element of consciousness that's missing when no sense is conveyed by what one sees or hears.

So within the totality of phenomenal experience we can distinguish more or less abstract elements, from basic sensory experiences like the experience of *red-here-now*, through depth perception, object gestalts, feelings of facial familiarity, to highly abstract language-based understanding experiences. We suggest there is a natural structural feature of the brain that the connectionist vehicle theory of consciousness can employ to account for this feature of experience. What we know of neural architecture indicates that the networks of which the brain is composed form a rough hierarchy. Some are very close to the sensory transducers, and receive their principal input from these, others are second-order (i.e., they receive their input from the first layer of networks), and so on. It is natural to suppose, on the connectionist vehicle theory of consciousness, that less abstract elements of experience correspond to stable patterns of activation in lower-order networks, while more abstract elements of experience correspond to stable patterns of activation in higher-order networks. Understanding experiences, in particular, (which incorporate both metacognitive and propositional forms of awareness) presumably correspond to stable patterns of activation in very high-order networks, networks that receive input from many sources, and are thus least modality specific and most subject to voluntary control. Thus, the continuum of degrees of abstractness evident in experience is explicable in terms of an underlying physical property of the brain – the hierarchical organization of its constituent networks. Again we find that a significant feature of phenomenal experience emerges naturally from the connectionist vehicle theory of consciousness.

---

[28] We know this, in part, because of the existence of *prosopagnosia*: an inability to recognize familiar faces. This deficit occurs as a result of characteristic kinds of lesions on the underside of the temporal and occipital lobes. Prosopagnosics are generally unable to recognize close family members by sight (although they can use other perceptual clues, such as voice quality, to identify them). One victim was even unfamiliar with his own face. In answer to the question "Are you able to recognize yourself in a mirror?", he replied, "Well, I can certainly see a face, with eyes, nose and mouth etc., but somehow it's not familiar; it really could be anybody" (reported in Zeki 1993, p.327). Thus, the feeling of facial familiarity is distinct from the experience of a face as an organized whole, or of its various identifiable components.

5.3 The Unity of Consciousness

Despite the compelling support for the connectionist vehicle theory that we've just rehearsed, this account will strike many as preposterous, given that, prima facie, it is at odds with some conventional wisdom concerning the *unity* of consciousness. Unity has traditionally been understood in terms of "oneness". To take a few representative examples: Baars describes conscious experience as "one thing after another…" (1988, p.83); Penrose says that "a characteristic feature of conscious thought…is its 'oneness' – as opposed to a great many independent activities going on at once" (1989, pp.398-399); and Paul Churchland tells us that "consciousness harbors the contents of the several basic sensory modalities within a *single unified experience*" (1995, p.214, emphasis in the original).[29] In other words, phenomenal experience, despite being polymodal, is unitary; a single thing.

But if consciousness, is just *one thing*, then there must be *one thing* that underlies it. Since it is implausible to suppose that the various distinct contents of instantaneous consciousness are encoded in a single representational vehicle, this suggests the need for a *single* consciousness-making mechanism or system of some kind. This is exactly what a number of theorists have proposed. Churchland, for example, develops the conjecture that phenomenal experience is the preserve of a particular neuroanatomical structure in the brain: the intralaminar nucleus in the thalamus. This structure has axonal projections to all areas of the cerebral hemispheres, and receives projections from those same areas. The brain thus contains a "grand informational loop" that "embraces all of the cerebral cortex", and which "has a bottleneck in the intralaminar nucleus" (p.215). Churchland claims (albeit tentatively – see p.223) that "a cognitive representation is an element of your current consciousness if, but only if, it is a representation…within the broad recurrent system [of the intralaminar nucleus]" (p.223). This conjecture allows him to account for the fact that "there are several distinct senses but only one unified consciousness" (p.214). What is crucial to this account is the existence of brain structures that act as a conduit – a functional bottleneck – through which information must pass in order to become conscious (the thalamic projection system and associated structures). These brain structures realize an *executive system* which is, in effect, a *single consciousness-making mechanism*.[30]

Clearly, when it comes to explaining the unity of consciousness, this avenue is not open to an advocate of the connectionist vehicle theory of phenomenal experience. The latter suggests that the neural basis of consciousness is both manifold and distributed. That is, it treats consciousness as a sum of independent phenomenal elements, each of which is generated at a different site in the cortex. Hence, on this account, what underlies consciousness is not one thing, but many. We might refer to this as a *multi-track* model of consciousness, by analogy with the recording technology that enables music to be distributed across numerous physically distinct tracks of a tape. Each consciousness-making mechanism in the cortex is like a separate recording track. Churchland's model, by contrast, is *single-track*. In a single-track recording there is no way to separate out the individual contributions of the musicians – they are packaged into a single structure. Likewise, in Churchland's model all of the different contentful elements are packaged together within a "single unified experience" (1995, p.214). On the face of it a multi-track model renders the unity of consciousness somewhat mysterious. A single-track model, on the other hand, is in the business of rendering consciousness unitary.

---

[29] Theorists who assert this don't necessarily take instantaneous consciousness to be restricted to a single modality. Paul Churchland, for example, regards our "single unified experience" as polymodal in character (see his 1995, pp.214-22).

[30] This account is strikingly similar to Baars' "Global Workspace" model of consciousness which we described earlier (see Section 1). Both Churchland and Baars take the unity of consciousness to be one of their principal explananda; both give informational feedback a pivotal role in their accounts of consciousness; and both identify the thalamic projection system and associated structures as potential realizers of this role.

However, it is pertinent, at this point, to note an ambiguity in the notion of unity. To assert that consciousness is unified is not necessarily to assert that it is literally a single entity, and thus must depend on a single neural vehicle or mechanism. Unity may also be construed in terms of *connectedness* and *coherence*. This property of consciousness is manifest both in the consonance displayed by the representational contents of the various modalities, and in the binding of phenomenal elements *within* modalities. And if this is the sense in which consciousness is unified then it's quite possible that the connectionist vehicle theory of consciousness is not so at odds with unity after all. In what follows we will offer an account of the coherence of consciousness that is consistent with the connectionist vehicle theory of consciousness. In order to do so it will first be necessary to unpack the notion of coherence a little further.

Phenomenal experience exhibits both intra-modal and inter-modal coherence. In our daily experience we sometimes only have one source of information regarding external objects: we hear the bird, but we can't see it; we see the ball (on the roof), but we can't feel it. In these cases we don't expect our various modes of experience to be in complete accord; their objects, being distinct, have no obligation to be in temporal or spatial register. However, very often we have access to information regarding a single object via two or more senses. When it comes to our own bodies, in particular, we are information rich. Thus, as one types on a keyboard the sound of one's fingers striking the keys is in synchrony with both the visual and tactile experiences of the key-strikes; the location of these same key-strikes, as revealed in visual experience, is compatible with their apparent auditory location; and one's proprioceptive and visual experiences of hand position are consonant. Inter-modal coherence is pervasive when our senses report on common events or objects. Within modalities we also discover a great deal of harmony among the distinct elements of experience. Vision, for example, provides us with information about color, shape, depth, and motion. But this information is not free-floating, it comes bound together in coherent phenomenal objects whose visual properties co-vary in a consistent fashion.

It is important to recognize that there are two aspects to coherence: 1) *temporal coherence*, as exemplified, for example, in the coincidence of visual, auditory and tactile experiences of a key-strike; and 2) *spatial coherence*, which manifests itself in numerous ways, e.g., we see our bodily parts in positions we feel them, we hear sounds emanating from objects in the direction we see them, we experience colors as confined to the boundaries of their objects, and so on. An approach to the unity of consciousness that is consistent with a multi-track model of consciousness emerges when we treat these two aspects of coherence separately. To begin with, it is not implausible to suppose that when phenomenal properties coincide temporally, either within modalities or across modalities, this is entirely due to the simultaneity of their vehicles (this suggestion is not new; see, e.g., Edelman 1989). So when a felt key-strike is temporally aligned with its seen counter-part in experience, we simply propose to explain this in terms of a brain architecture that generates simultaneous vehicles in those two modalities. It's reasonable to believe that evolutionary pressures will have conspired to wire the brain in this way, given the tight temporal constraints that attend useful interaction with our local environment.[31]

Clearly, simultaneity of vehicles is not going to have much bearing on spatial coherence. For when we seek to explain this form of coherence we must contend with what Akins refers to as the *Spatial Binding Problem*, viz.: "given that the visual system processes different properties of the stimulus at spatially distinct sites, how is it possible that we perceive the world in the spatially coherent manner that we do?" (1996, p.30). Single-track theories of consciousness take this problem in their stride by refusing to identify visual experience solely with the machinations

---

[31] What we're suggesting here is that in order that we be able to respond appropriately to rapidly changing local conditions, the various determinants of a behavioural response (visual input, tactile input, proprioceptive input, and so forth) will need to be brought to bear roughly synchronously, in order that they not interfere with each other. Thus, the vehicles of these various kinds of information are likely to *be* synchronous (as a result of selective pressures on brain wiring). See also Churchland & Sejnowski 1992, p.51.

of the visual system. A visual content does not become conscious until it enters the consciousness-making system to which all conscious information is subject. But a multi-track theorist is in the business of identifying experience with the neural vehicles of explicit information, so the binding problem is pressing. However, it's not clear that this problem is intractable from the perspective of the connectionist vehicle theory of consciousness. Indeed, it may be no more than a pseudo-problem generated by adopting what Akins calls the *Naive Theory of Perception*: "the thesis that properties of the world must be represented by 'like' properties in the brain, and that these representations, in turn, give rise to phenomenological experiences with similar characteristics" (1996, p.14). In relation to, say, spatial properties, this theory requires that the spatial coherence of visual information "must be mimicked by the spatial unity of the representational vehicles themselves" (p.31). And this surely is a *naive* theory. We don't expect the green of grass to be represented by green-colored neural vehicles. Why, therefore, should we expect spatial properties of the world to be represented by corresponding spatial properties of the brain? So long as the contributing sensory systems represent their common object *as* located in the one place, then the experience of object location ought to be both inter-modally and intra-modally coherent. In particular, the only intra-modal "binding" we can reasonably expect is a binding at the level of contents. In order for the various properties of, say, a visual object to be experienced as unified, the visual system need only represent them *as* occurring in a common region of space. (This implies, of course, that each element of visual experience, for example, in addition to its non-spatial content, also incorporates spatial information. That is, the basic elements of vision are *color-x-at-location-y*, and so on.) And to deal with multiple, co-occurrent objects we simply need to posit a number of such "content-bindings" realized by multiple, simultaneous representational vehicles.

We haven't yet touched on a further important way in which consciousness is unified. There is a real sense in which 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 – they are yours and no-one else's. It is perhaps this salient dimension that Churchland is really alluding to when he talks in terms of consciousness harbouring "the contents of the several basic sensory modalities within a *single unified experience*" (1995, p.214); but, *pace* Churchland, it is not the *experience* that is unified; the unification is at the level of the cognitive subject: the various phenomenal elements, issuing from the different sensory faculties, all "belong to" or in some sense "constitute" the one subject. We will call this form of unity *subject unity*. Given the multi-track nature of our account of consciousness, there is an issue as to how our sense of subject unity arises.

There are at least two ways of explaining subject unity consistent with our vehicle theory. On the one hand, we can treat it as that very abstract sense of self that arises out of our ongoing personal narrative, the story we tell about ourselves, and to ourselves, practically every waking moment. This narrative, a product of those centres responsible for natural language comprehension and production, comprises a serial stream of self-directed thought (one that non-language using animals presumably lack). On the other hand, we can explain one's feeling of subject unity in terms of the *confluence* of the points of view generated by the individual phenomenal elements that make up our instantaneous conscious experience. While these phenomenal elements arise independently in every mode of experience, each of them encompasses a space with a privileged locus, a point with respect to which every content is "projected". Consequently, so long as the various modalities represent their respective kinds of information *as* located with respect to the *same* projective locus, this will generate a single phenomenal subject located at a particular point in space.[32] Rejection of the Naive Theory of

[32] It is important to be clear that this solution explains one's *feeling* of subject unity (a first-person fact); it is does not, nor is it required to, explain how your experiences are, ontologically speaking, different from, e.g., your neighbor's (a third-person fact). This latter is simply explained by the fact that your experiences are identical with explicit vehicles in your head, which are *physically distinct* from the vehicles found in your neighbor's head.

Perception, in particular, rejection of the view that the representation of spatial properties necessarily involves corresponding spatial properties in the brain, undermines the idea that such a common point of view must necessarily involve a single-consciousness making mechanism.

5.4 The Explanatory Gap

The connectionist vehicle theory we are advocating identifies phenomenal experiences with the stable patterns of activation generated in the brain's neural networks. But some will find this suggestion objectionable for the reason that it doesn't seem to provide us with a *satisfying* reductive explanation of consciousness. A reductive explanation is satisfying when there is a "perspicuous nexus" between the postulated micro-mechanism and the macro-phenomenon in question, such that we can "see" the connection between them. (Cottrell 1995). We are happy identifying water with H_2O, to use the standard example, because we understand how the molecular properties of the latter must give rise to the familiar properties of the former. But it is precisely this kind of intelligible connection that appears to be lacking in the case of our proposal: what is it *about* stable activation patterns, one might ask, that they should give rise to the familiar differential properties of phenomenal consciousness?

This, of course, raises the special explanatory difficulties associated with phenomenal consciousness. Quite independent of finding a robust neural correlate of phenomenal experience, is the problem of explaining how *any* kind of physical object could possess this remarkable property. This is the so-called "hard" problem of consciousness (Nagel 1974; Chalmers 1995, 1996), which creates an "explanatory gap" between our materialist hypotheses about the neural substrate of consciousness and its phenomenal properties (Levine 1983, 1993). The problem, in a nutshell, is that whatever physical or functional property of the brain we cite in our attempt to explain consciousness, we can always conceive of a creature instantiating this property *without* being subject to phenomenal experiences. Consequently, *any* materialist theory of consciousness tends to have an air of impotence about it.

The *least* we can say of the connectionist vehicle theory of consciousness, is that it's no worse off, in this respect, than any other current theory. But there is more we can say in this regard. What we can properly conceive is not fixed, but narrows with the development of our scientific understanding. What today we think is unimaginable might tomorrow merely be indicative of the fact that we possessed insufficient information. To borrow an example of Cottrell's, anybody lacking a knowledge of special relativity will think that it is conceivable that some particles might travel faster than photons: "One imagines the photon as a tiny bullet, speeding along; and one imagines some bullet x overtaking it. But once we know a little about relativity, we begin to see that this imagining is not really coherent; if we are pushed into confronting the implications of x's overtaking the photon, we will see that it leads to absurdities" (1995, p.99). The same point can be applied to our understanding of consciousness. The more we learn about the connection between the brain's neural substrate and phenomenal consciousness, the "more we have in the way of explanatory hooks on which to hang something that could potentially close the explanatory gap" (Block 1995, p.245, fn.5 – see also Flanagan 1992, p.59; and Van Gulick 1993). In particular, if we can find a neural mechanism that mirrors in a systematic fashion the complex structural properties of phenomenal experience, it may begin to seem natural that this mechanism must give rise to consciousness. The explicit representation of information in neurally realized PDP networks, we think, is just such a mechanism, and hence this connectionist vehicle theory has the potential to go some way towards bridging the explanatory gap (see also Lloyd 1996).

We have already seen, in Section 5.2, how this connectionist hypothesis accounts for many of the structural and temporal properties of our instantaneous experience. But what might not be so readily apparent is that it can provide a systematic account of the *similarities* and *differences* between the phenomenal elements that comprise this complex.

Consider, for example, our perception of *color*. Human beings are capable of discriminating at least 10,000 distinct colors, organized in a fine-grained "color metric" which enables us to say whether one color is more similar to a second than to a third; whether a color is between two other colors; and so forth (Hardin 1988). As is well known, connectionist activation pattern representation provides a powerful explanation of such a metric (see, e.g., Churchland 1995; Churchland & Sejnowski 1992, Chp.4; Clark 1993; and Rumelhart & McClelland 1986, Chps.1-3). The spiking activity across a neural network can be represented in terms of a hyper-dimensional *activation space*, the points in which describe individual activation patterns. And the *geometrical* properties of this activation space, which model the structural relations between the activation patterns realizable in the network, can be invoked to explain the *phenomenal* relations that obtain between conscious experiences in any one domain. Color experiences that are very different (say the experience of red versus green), for instance, can be thought to correspond with stable patterns of activation that map onto widely separated points in this activation space; while points that are near neighbours in this space correspond with color experiences that are phenomenally similar.

This is striking enough. But even more striking is the fact that this same connectionist approach to consciousness provides the beginnings of an explanatory framework that can account for how the one neural substrate is capable of generating all the different *kinds* of experience (both within and across sensory modalities) we're capable of entertaining. Remaining for the moment with visual phenomenology, we clearly need a neural mechanism that can do more than explain the phenomenal differences between colors; it must also be capable of accounting for the differences between color experiences, and size, shape, texture and motion experiences. And once we look *across* modalities, the differences become even more dramatic. This neural mechanism must be capable of explaining the differences between colors, sounds, tastes, smells, and so forth. It must also have the resources to account for the differences between more concrete and more abstract experiences in each of these modalities. And it must be able to distinguish between the various kinds of linguistically-mediated experiences in a systematic fashion.

But all these differences are explicable, we think, with the resources of connectionist activation pattern representations. Of course, the difference between, say, an experience of red and the sound of a trumpet cannot be explained by recourse to different *points* in the one activation space. Rather, to explain the similarities and differences between *kinds* of experience, one appeals to the similarities and differences *between activation spaces*. Activation spaces differ according to both "dimensionality", which is determined by the number of neurons contained in a neural network, and "shape", with depends on precisely how these neurons are connected. Both of these features can be brought to bear in accounting for the differences between broad classes of experience. What *unites* color experiences is that they correspond to patterns of activation in an activation space with a particular geometric structure (shape and dimensionality). But equally, what *distinguishes* them from experiences of, say, sound, are these same geometric properties, properties which distinguish one neural network from another, and hence, on our account, one *kind* of phenomenology from another.

Of course, there is a great deal of explanatory work to be done here in linking these different activation spaces in a systematic fashion to their proprietary representational domains. This is a task for a theory of mental content; a theory that can explain how the different activation pattern representations realizable in a particular activation space actually receive their distinct semantic interpretations.[33] But precisely because this connectionist vehicle theory has

---

[33] One natural suggestion in this regard, though one that is not very popular in the contemporary philosophy of mind, is that this linkage, at least for some representational states, might be unpacked in terms of *structural isomorphisms* that obtain between stable activation patterns and the objects in the represented domain (see, e.g., Cummins 1996, Chp.7; Gardenfors 1996; Palmer 1978; and Swoyer 1991).

the resources to model all of the similarities and differences between these representational states, it does have the potential to close the explanatory gap.

6 Conclusion

In this paper we've done something that is singularly unpopular in contemporary cognitive science: we've developed and defended a vehicle theory of phenomenal consciousness; that is, a theory that identifies phenomenal experience with the vehicles of explicit representation in the brain. Such a position is unpopular, we think, not in virtue of the inherent implausibility of vehicle theories, but largely because of the influence (both explicit and implicit) exerted by the classical computational theory of mind. With the advent of connectionism it is time to take a fresh look at these issues. This is because connectionism provides us with a different account of both information coding and information processing in the brain, especially with respect to the role of inexplicitly coded information, and hence opens up new regions of the theoretical landscape for serious exploration. Given the many difficulties connected with existing computational theories of consciousness, this is surely to be welcomed.

The connectionist vehicle theory of phenomenal experience forces us to re-assess some common wisdom about consciousness. It suggests that instantaneous consciousness is not a single, monolithic state, but a complex amalgam of distinct and relatively independent phenomenal elements. Consequently, it also suggests that our ongoing consciousness is not a single stream, but a mass of tributaries running in parallel. And it suggests that we are conscious of a good deal more information at any one moment in time than theorists have traditionally supposed (Lloyd 1991, pp.454-5 makes a similar point). But each of these revisions to the standard lore on consciousness is defensible on independent grounds, as our examination of both the phenomenological and neuroscientific evidence demonstrates. Consciousness, we have seen, is a rich tapestry woven from many threads. And hence the connectionist vehicle theory we have been promoting, with its multiplicity of consciousness-making mechanisms scattered right across the brain, is precisely the sort of account we need.

Beyond these incentives, we feel that our connectionist account of consciousness is ideally pitched for cognitive science. By tying phenomenal experience to the explicit representation of information, and hence finding a place for consciousness at the foundation of the brain's information processing capacity, this thesis provides the discipline with a *principled* computational theory of phenomenal consciousness. Phenomenal consciousness is not an emergent product of complex information processing, nor of sufficiently rich and widespread information processing relations; rather, consciousness is the mechanism whereby information is explicitly encoded in the brain, and hence is a fundamental feature of cognition.

Acknowledgments

We would like to thank Derek Browne, Rich Carlson, George Couvalis, Greg Currie, Don Dulany, Denise Gamble, Jon Jureidini, Dan Lloyd, Greg O'Hair, Bruce Mangan, Drew McDermott, Chris Mortensen, Ian Ravenscroft, John Sutton and a number of anonymous referees of this journal for their very helpful comments on earlier versions of this paper. We are also grateful to many audiences at talks on this material for their criticisms.

References

Akins, K. (1996) Lost the Plot? Reconstructing Dennett's Multiple Drafts Theory of Consciousness. Mind and Language 11: 1-43.

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

Bechtel, W. (1988a) Connectionism and rules and representation systems: Are they compatible? Philosophical Psychology.1: 1-15.

Bechtel, W. (1988b) Connectionism and interlevel relations. Behavioral and Brain Sciences 11: 24-25.

Bechtel, W. & Abrahamsen, A. (1991) Connectionism and the Mind. Blackwell.

Bisiach, E. (1992) Understanding consciousness: Clues from unilateral neglect and related disorders. In Milner and Rugg, 1992.

Block, N. (1993) Book review of Dennett's *Consciousness Explained*. Journal of Philosophy 90:181-93.

Block, N. (1995) On a confusion about a function of consciousness. Behavioral and Brain Sciences 18:227-87.

Cam, P. (1984) Consciousness and content-formation. Inquiry 27: 381-97.

Campion, J., Latto, R. & Smith, Y.M. (1983) Is blindsight an effect of scattered light, spared cortex, and near-threshold vision?. Behavioral and Brain Sciences 6:423-86

Chalmers, D. (1995) Facing up to the problem of consciousness. Journal of Consciousness Studies 2: 200-19.

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

Charland, L.C. (1995) Emotion as a natural kind: Towards a computational foundation for emotion theory. Philosophical Psychology 8: 59-84.

Chomsky, N. (1980) Rules and representations. Behavioral and Brain Sciences 3: 1-62.

Churchland, P.M. (1995) The Engine of Reason, the Seat of the Soul. MIT Press.

Churchland, P.S. & Sejnowski, T (1992) The Computational Brain. MIT Press.

Clark, A. (1989) Microcognition: Philosophy, Cognitive Science, and Parallel Distributed Processing. MIT Press

Clark, A. (1993) Associative Engines: Connectionism, Concepts and Representational Change. MIT Press.

Clark, A. & Karmiloff-Smith, A. (1993) The cognizer's innards: A psychological and philosophical perspective on the development of thought. Mind and Language 8: 487-519.

Clark, A. & Thornton, C. (forthcoming) Trading spaces: Computation, representation and the limits of uninformed learning. Behavioral and Brain Sciences.

Cleland, C.E. (1993) Is the Church-Turing thesis true?. Minds and Machines 3: 283-313.

Corteen, R.S. (1986) Electrodermal responses to words in an irrelevant message: A partial reappraisal. Behavioral and Brain Sciences 9: 27-28.

Corteen, R.S. & Wood, B. (1972) Electrodermal responses to shock-associated words in an unattended channel. Journal of Experimental Psychology 94: 308-13.

Cottrell, A. (1995) *Tertium datur?* Reflections on Owen Flanagan's *Consciousness Reconsidered*. Philosophical Psychology 8: 85-103.

Cummins, R. (1986) Inexplicit representation. In: The Representation of Knowledge and Belief, ed. M.Brand & R.Harnish. University of Arizona Press.

Cummins, R. (1996) Representations, Targets, and Attitudes. MIT Press.

Cummins, R. & Schwarz, G. (1991) Connectionism, computation, and cognition. In: Connectionism and the Philosophy of Mind, ed. T.Horgan & J.Tienson. Kluwer.

Cussins, A. (1990) The connectionist construction of concepts. In: The Philosophy of Artificial Intelligence, ed. M.Boden. Oxford University Press

Crick, F. (1984) Function of the thalamic reticular complex: The searchlight hypothesis. Proceedings of the National Academy of Sciences, USA 81: 4586-90.

Dennett, D.C. (1982) Styles of mental representation. Proceedings of the Aristotelian Society New Series 83: 213-26.

Dennett, D.C. (1984) Cognitive wheels: The frame problem of AI. In Minds, Machines and Evolution, ed. C.Hookway. Cambridge University Press.

Dennett, D.C. (1991) Consciousness Explained. Little Brown.

Dennett, D.C. (1993) The message is: There is no *me*dium. Philosophy and Phenomenological Research 53: 919-31.

Dennett, D.C & Kinsbourne, M. (1992) Time and the observer: The where and when of consciousness in the brain. Behavioral and Brain Sciences 15:183-247.

Dienes, Z., Broadbent, D.E. & Berry, D. (1991) Implicit and explicit knowledge bases in artificial grammar learning. Journal of Experimental Psychology: Learning, Memory, and Cognition 17: 875-8.

Dietrich, E. (1989) Semantics and the computational paradigm in cognitive psychology. Synthese 79: 119-41.

Dretske, F. (1993) Conscious experience. Mind 102: 263-83.

Dretske, F. (1995) Naturalizing the Mind. MIT Press.

Dulany, D.E. (1991) Conscious representation and thought systems. In: Advances in Social Cognition IV, ed. R.S.Wyer Jr. & T.K.Srull. Lawrence Erlbaum

Dulany, D.E. (1996) Consciousness in the explicit (deliberative) and implicit (evocative). In: Scientific Approaches to Consciousness, ed, J.Cohen & J.Schooler. Lawrence Erlbaum

Dulany, D.E., Carlson, R.A. & Dewey, G.I.(1984) A case of syntactical learning and judgement: How conscious and how abstract? Journal of Experimental Psychology: General 113: 541-55.

Edelman, G.M. (1989) The Remembered Present: A Biological Theory of Consciousness. Basic Books.

Field, H. (1978) Mental representation. Erkenntnis 13: 9-61.

Flanagan, O. (1992) Consciousness Reconsidered. MIT Press.

Fodor, J.A. (1975) The Language of Thought. MIT Press.

Fodor, J.A. (1981) Representations. MIT Press.

Fodor, J.A. (1983) The Modularity of Mind. MIT Press.

Fodor, J.A. (1987) Psychosemantics. MIT Press.

Fodor, J.A. & Pylyshyn, Z.W. (1988) Connectionism and cognitive architecture: A critical analysis. Cognition 28: 3-71.

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

Hardin, C.L. (1988) Color for Philosophers. Hackett.

Harman, G. (1973) Thought. Princeton University Press

Hatfield, G. (1991) Representation in perception and cognition: Connectionist affordances. In Philosophy and Connectionist Theory, ed. W.Ramsey, S.Stich & D.Rumelhart. Lawrence Erlbaum.

Haugeland, J. (1981). Semantic engines: An introduction to mind design. In: Mind Design, ed. J.Haugeland. MIT Press.

Haugeland, J. (1985) Artificial Intelligence: The Very Idea. MIT Press.

Holender, D. (1986) Semantic activation without conscious awareness in dichotic listening, parafoveal vision, and visual masking: A survey and appraisal. Behavioral and Brain Sciences 9: 1-66.

Hopcroft, J.E. & Ullman, J.D. (1979) Introduction to Automata Theory, Languages and Computation. Addison Wesley Publishing Company

Horgan, T. & Tienson, J. (1989) Representations without rules. Philosophical Topics 27: 147-74.

Jackendoff, R. (1987) Consciousness and the Computational Mind. MIT Press.

James, W. (1890) The Principles of Psychology. Holt.

Johnson-Laird, P.N. (1988) The Computer and the Mind: An Introduction to Cognitive Science. Fontana Press.

Johnston, W.A. & Dark, V.J. (1982) In defense of intraperceptual theories of attention. Journal of Experimental Psychology: Human Perception and Performance 8: 407-21.

Johnston, W.A. & Wilson, J. (1980) Perceptual processing of nontargets in an attention task. Memory and Cognition 8: 372-7

Kinsbourne, M. (1988) Integrated field theory of consciousness. In: Consciousness in Contemporary Science, ed. A.Marcel & E.Bisiach. Clarendon Press.

Kinsbourne, M. (1995) Models of consciousness: Serial or parallel in the brain? In: The Cognitive Neurosciences ed. M.Gazzaniga. MIT Press.

Lackner, J.R. & Garrett, M.F. (1972) Resolving ambiguity: Effects of biasing context in the unattended ear. Cognition 1: 359-72.

Levine, J. (1983) Materialism and qualia: The explanatory gap. Pacific Philosophical Quarterly 64: 354-61.

Levine, J. (1993) On leaving out what it is like. In: Consciousness: Psychological and Philosophical Essays, ed. M.Davies & G.Humphreys. Blackwell.

Lloyd, D. (1988) Connectionism in the golden age of cognitive science. Behavioral and Brain Sciences 11: 42-43.

Lloyd, D. (1991) Leaping to conclusions: Connectionism, consciousness, and the computational mind. In: Connectionism and the Philosophy of Mind, ed.T.Horgan & J.Tienson. Kluwer.

Lloyd, D. (1995) Consciousness: A connectionist manifesto. Minds and Machines 5: 161-85.

Lloyd, D. (1996) Consciousness, connectionism, and cognitive neuroscience: A meeting of the minds. Philosophical Psychology 9: 61-79.

Mandler, G. (1985) Cognitive Psychology: An Essay in Cognitive Science. Lawrence Erlbaum.

Mangan, B. (1993a) Dennett, consciousness, and the sorrows of functionalism. Consciousness and Cognition 2: 1-17.

Mangan, B. (1993b) Taking phenomenology seriously: The "fringe" and its implications for cognitive research. Consciousness and Cognition 2: 89-108.

Mangan, B. (1996) Against functionalism: Consciousness as an information-bearing medium. Presented at the Tucson II conference on consciousness, Arizona

Marcel, A.J. (1983) Conscious and unconscious perception: Experiments on visual masking and word recognition. Cognitive Psychology 15: 197-237.

MacKay, D.G. (1973) Aspects of a theory of comprehension, memory and attention. Quarterly Journal of Experimental Psychology 25:22-40

McClelland, J.L. & Rumelhart, D.E., eds. (1986) Parallel Distributed Processing: Explorations in the Microstructure of Cognition Vol. 2: Psychological and Biological Models. MIT Press.

Milner, A. & Rugg, M., eds. (1992) The Neuropsychology of Consciousness. Academic Press.

Nagel, T. (1974). What is it like to be a bat? Philosophical Review 83: 435-50.

Nelson, T.O. (1978) Detecting small amounts of information in memory: Savings for nonrecognized items. Journal of Experimental Psychology: Human Learning and Memory 4: 453-68

Newman, J. (1995) Thalamic contributions to attention and consciousness. Consciousness and Cognition 4: 172-93.

Newell, A. (1980) Physical symbol systems. Cognitive Science 4: 135-83.

Newstead, S.E. & Dennis, I. (1979) Lexical and grammatical processing of unshadowed messages: A reexamination of the MacKay effect. Quarterly Journal of Experimental Psychology 31: 477-88

Nolan, K.A. & Caramazza, A. (1982) Unconscious perception of meaning: A failure to replicate. Bulletin of the Psychonomic Society 20: 23-6.

O'Brien, G. (1993) The connectionist vindication of folk psychology. In: Folk Psychology and the Philosophy of Mind, ed. S.Christensen & D.Turner. Lawrence Erlbaum.

Palmer, S. (1978) Fundamental aspects of cognitive representation. In: Cognition and Categorization, ed. E.Rosch & B.Lloyd. Lawrence Erlbaum.

Penrose, R. (1989) The Emperor's New Mind. Penguin Books

Perenin, M.T. (1978) Visual function within the hemianopic field following early cerebral hemidecortication in man. II. Pattern discrimination. Neuropsychologia 16: 697-708.

Perenin, M.T. & Jeannerod, M. (1975) Residual vision in cortically blind hemifields. Neuropsychologia 13: 1-7.

Perruchet, P. & Pacteau, C. (1990) Synthetic grammar learning: Implicit rule abstraction or explicit fragmentary knowledge? Journal of Experimental Psychology: General 119: 264-75.

Purcell, D.G., Stewart, A.L. & Stanovich, K.K. (1983) Another look at semantic priming without awareness. Perception and Psychophysics 34: 65-71.

Pylyshyn, Z.W. (1980) Computation and cognition: Issues in the foundations of cognitive science. Behavioral and Brain Sciences 3: 111-69.

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

Pylyshyn, Z.W. (1989) Computing in cognitive science. In: Foundations of Cognitive Science, ed. M.Posner. MIT Press

Ramsey, W., Stich, S. & Rumelhart, D.E., eds. (1991) Philosophy and Connectionist Theory. Lawrence Erlbaum.

Reber, A.S. (1967) Implicit learning of artificial grammars. Journal of Verbal Learning and Verbal Behavior 5: 855-63.

Rey, G. (1992) Sensational sentences. In: Consciousness: Psychological and Philosophical Essays, ed. M.Davies & G.Humphreys. Blackwell.

Rubel, L.A. (1989) Digital simulation of analog computation and Church's thesis. Journal of Symbolic Logic 54: 1011-7.

Rumelhart, D.E. (1989) The architecture of mind: A connectionist approach. In: Foundations of Cognitive Science, ed. M. Posner. MIT Press.

Rumelhart, D.E. & McClelland, J.L., eds. (1986) Parallel Distributed Processing: Explorations in the Microstructure of Cognition Vol 1: Foundations. MIT Press.

Rumelhart, D.E., Smolensky, P., McClelland, J.L. & Hinton, G.E. (1986) Schemata and sequential thought processes in PDP models. In McClelland & Rumelhart 1986.

Sacks, O. (1985) The Man Who Mistook His Wife For a Hat. Picador.

Schacter, D. (1989) On the relation between memory and consciousness: Dissociable interactions and conscious experience. In: Varieties of Memory and Consciousness: Essays in Honour of Endel Tulving, ed. H.Roediger & F.Craik. Erlbaum.

Schwartz, N. (1990) Feelings and information: Informational and motivational functions of affective states. In: Handbook of Motivation and Cognition: Foundations of Social Behaviour, ed. R.Sorrentino & E.Higgins. Guilford Press.

Searle, J.R. (1983) Intentionality. Cambridge University Press

Sejnowski, T.J. (1986) Open questions about computation in cerebral cortex. In: McClelland & Rumelhart 1986.

Sejnowski, T.J. & Rosenberg, C. (1987) Paralled networks that learn to pronounce English text. Complex Systems 1: 145-68.

Shallice, T. (1988a) From Neuropsychology to Mental Structure. Cambridge University Press.

Shallice, T. (1988b) Information-processing models of consciousness: Possibilities and problems. In: Consciousness in Contemporary Science, ed. A.Marcel & E.Bisiach. Clarendon Press.

Shanks, D.R. & St. John, M.F. (1994) Characteristics of dissociable human learning systems. Behavioral and Brain Sciences 17:367-447

Smolensky, P. (1988) On the proper treatment of connectionism. Behavioral and Brain Sciences 11: 1-23.

Smolensky, P. (1987). The constituent structure of connectionist mental states: A reply to Fodor and Pylyshyn. Southern Journal of Philosophy 26(Supplement): 137-61.

Swoyer, C. (1991) Structural representation and surrogative reasoning. Synthese, 87: 449-508.

Sterelny, K. (1990) The Representational Theory of Mind. Blackwell

Strawson, G. (1994) Mental Reality. MIT Press

Tienson, J.L. (1987) An introduction to connectionism. The Southern Journal of Philosophy 26(Supplement): 1-16.

Tye, M. (1992) Visual qualia and visual content. In: The Contents of Experience, ed. T.Crane. Cambridge University Press.

Tye, M. (1996) The function of consciousness. Nous 30: 287-305.

Tye, M. (forthcoming) A representational theory of pains and their phenomenal character. In: Essays on Consciousness, ed. N.Block, O.Flanagan, & G.Guveldere. MIT Press.

Umilta, C. (1988) The control operations of consciousness. In: Consciousness in Contemporary Science, ed. A.Marcel & E.Bisiach. Clarendon Press.

Van Gelder, T. (1990) Compositionality: A connectionist variation on a classical theme. Cognitive Science 14: 355-84

Van Gulick, R. (1993) Understanding the phenomenal mind: Are we all just armadillos? In: Consciousness: Psychological and Philosophical Essays, ed. M.Davies & G.Humphreys. Blackwell.

Von Eckardt, B. (1993) What is Cognitive Science? MIT Press.

Weiskrantz, L. (1980) Varieties of residual experience. Quarterly Journal of Experimental Psychology 32: 365-86.

Weiskrantz, L. (1986) Blindsight: A Case Study and Implications. Clarendon Press.

Weiskrantz, L., Warrington, E., Sanders, M. & Marshall, J. (1974) Visual capacity in the hemianopic field following a restricted occipital ablation. Brain 97: 709-28.

Zeki, S. (1993) A Vision of the Brain. Blackwell.